

第六章 统计量与抽样分布

- 随机样本与统计量
- χ^2 分布 t 分布 F 分布
- 正态总体下的抽样分布

- **数理统计** 是一门以数据为基础的学科, 可以定义为收集数据, 分析数据和由数据得出结论的一组概念、 原则和方法。
- 例如: 若规定灯泡寿命低于1000小时者为次品, 如何确定次品率? 由于灯泡寿命试验是破坏性试验, 不可能把整批灯泡逐一检测, 只能抽取一部分灯泡作为样本进行检验, 以样本的信息来推断总体的信息, 这是数理统计研究的问题。

6.1 随机样本与统计量

- 总体：研究对象的全体；
- 个体：总体中的成员；
- 总体的容量：总体中包含的个体数；
- 有限总体：容量有限的总体；
- 无限总体：容量无限的总体，通常将容量非常大的总体也按无限总体处理。

- 例：1) 了解某校“大学生的月消费水平”。总体是该校大学生全体。这是一个有限总体，每个大学生有许多指标，我们关注的是学生“过去6个月平均每月的花费”这一指标。

- 2) 了解某城市的空气质量情况，调查该城市的**PM2.5**值。这是一个无限总体，描述空气质量有许多指标，而我们仅关心**PM2.5**值。
- 3) 研究某种药物在人体中的吸收情况。这是一个有限总体，但数量非常巨大，我们常把它看出无限总体。

为了采用数理统计方法进行分析，首先要收集数据，数据收集方法一般有两种。

(1) 通过调查、记录收集数据。如为了调查大学生“过去**6**个月平均每月的花费”，可以进行问卷调查；要了解**PM2.5**值，需要在城市设立若干**PM2.5**监测站点，定时收集数据。

(2) 通过实验收集数据。如为了了解药物吸收情况，要征集若干志愿者，把他们分成若干组，观察他们服药后不同时间点药物含量数据。

关于调查数据和实验数据的收集可以根据数据本身的特点有多种不同的方法和设计，有专门的课程讲授，这里不作详细介绍。

- 研究对象的某个指标 X , 对于不同的个体来说有不同的取值, 这些取值构成一个分布, 因此 X 可以看成是一个随机变量. 有时候就把 X 称为总体. 假设 X 的分布函数为 $F(x)$, 也称 $F(x)$ 为总体.

- 数理统计主要任务是从总体中抽取一部分个体, 根据这部分个体的数据对总体分布给出推断. 被抽取的部分个体叫做总体的一个 样本.

▶ **随机样本：**从总体中随机地取 n 个个体, 称为一个随机样本。

▶ **简单随机样本：**满足以下两个条件的随机样本 (X_1, X_2, \dots, X_n) 称为容量是 n 的简单随机样本。

1. **代表性：**每个 X_i 与 X 同分布；

2. **独立性：** X_1, X_2, \dots, X_n 是相互独立的随机变量。

[说明]：后面提到的样本均指简单随机样本。

[注意]: 一个容量为 n 的样本 X_1, X_2, \dots, X_n

是指 n 个独立与总体分布相同的随机变量。

一旦对样本进行观察，得到实际数值

$$x_1, x_2, \dots, x_n$$

称为样本观察值（或样本值）。

两次观察，样本值可能是不同的。

■ 如何取得的样本才称是简单随机样本？

对于有限总体, 采用放回抽样就能得到简单随机样本.

但当总体容量很大的时候,放回抽样有时候很不方便, 因此在实际中当总体容量比较大时,通常将不放回抽样所得到的样本近似当作简单随机样本来处理.

对于无限总体, 一般采取不放回抽样.

例1.1 有4个学生参加《概率论与数理统计》课程考试，成绩分别为88, 75, 70, 63. 现从中抽取容量为2 的样本，列出全部的样本.

答：共有16个样本，分别为：

(88, 88)	(88, 75)	(88, 70)	(88, 63)
(75, 88)	(75, 75)	(75, 70)	(75, 63)
(70, 88)	(70, 75)	(70, 70)	(70, 63)
(63, 88)	(63, 75)	(63, 70)	(63, 63)

▶ **统计量：** 样本的不含任何未知参数的函数。

▶ **常用统计量：** 设 (X_1, X_2, \dots, X_n) 为取自总体 X 的简单随机样本。常用的统计量如下：

1. 样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
2. 样本方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, S 为样本标准差
3. 样本矩 k 阶矩: $A_k = \frac{1}{n} \sum_{i=1}^n X_i^k \quad (k=1, 2, \dots)$
 k 阶中心矩: $B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k \quad (k=1, 2, \dots)$

[思考题]:

对于总体 X , X_1, \dots, X_n 是来自总体的样本,
设下列数字特征存在,

$$E(X) = \mu, \text{Var}(X) = \sigma^2, E(X^k), E[(X - \mu)^k],$$

问: (1) \bar{X} 与 μ , (2) S^2 与 σ^2 ,

(3) A_k 与 $E(X^k)$, (4) B_k 与 $E[(X - \mu)^k]$

都相等吗?

答: 不对。前者是随机变量, 观察两次得到的统计量的值可能不一样;
后者是数, 可能已知也可能未知。

当总体数字特征未知时 (设各阶矩存在)

- 一般, 用样本均值 \bar{X} 作为总体均值 $\mu = E(X)$ 的估计;
- 用样本方差 S^2 作为总体方差 $\sigma^2 = E(X - \mu)^2$ 的估计;
- 用样本原点矩 A_k 作为总体原点矩 $\mu_k = E(X^k)$ 的估计;
- 用样本中心矩 B_k 作为总体中心矩 $\nu_k = E(X - \mu)^k$ 的估计.
- 总体方差的估计可以用 S^2 也可以用 B_2 , 主要的区别涉及到“无偏性” (这个概念将在第七章讨论).

例1.2 接例1.1，总体为88，75，70，63，显然，总体均值为74. 计算全部16个样本的样本均值.

样本 编号	样本	样本 均值	样本 编号	样本	样本 均值	样本 编号	样本	样本 均值
1	(88,88)	88	7	(75,70)	72.5	13	(63,88)	75.5
2	(88,75)	81.5	8	(75,63)	69	14	(63,75)	69
3	(88,70)	79	9	(70,88)	79	15	(63,70)	66.5
4	(88,63)	75.5	10	(70,75)	72.5	16	(63,63)	63
5	(75,88)	81.5	11	(70,70)	70	16个样本均值的平均 为74		
6	(75,75)	75	12	(70,63)	66.5			

从中看到，用样本均值估计总体均值，可能估计过高，可能估计过低。

所有样本均值的平均值恰好是总体均值。(无偏性)

6.2 χ^2 分布 t 分布 F 分布

(一) χ^2 分布

定义：设随机变量 X_1, X_2, \dots, X_n 相互独立，

$X_i \sim N(0, 1) \quad (i = 1, 2, \dots, n)$ 则称

$$\chi^2 = \sum_{i=1}^n X_i^2 \quad (1)$$

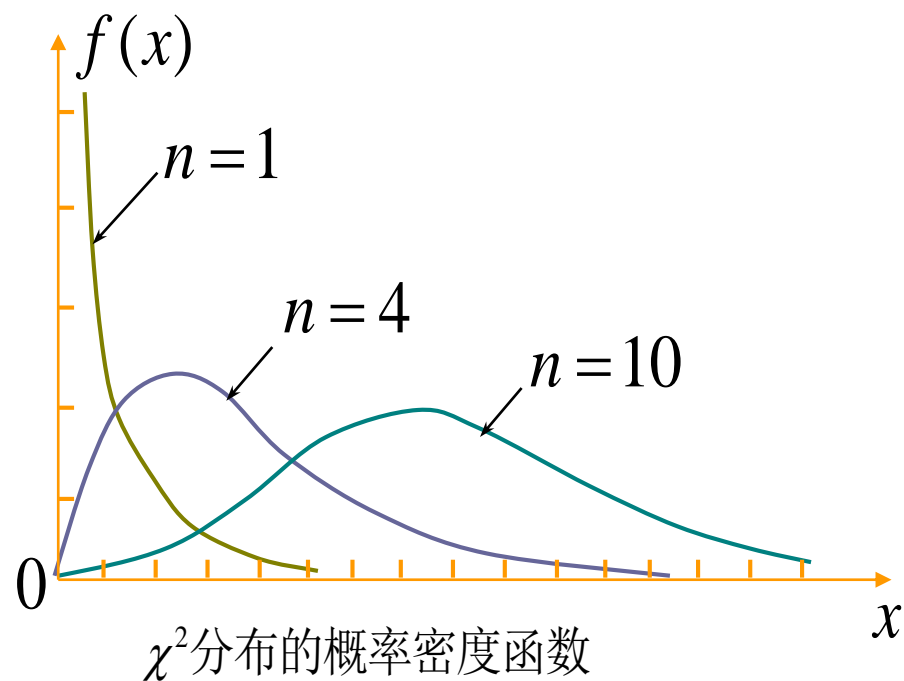
服从自由度为 n 的 χ^2 分布，记为 $\chi^2 \sim \chi^2(n)$

其中，自由度指(1)式右端包含的独立变量个数.

$\chi^2(n)$ 分布的概率密度函数为:

$$f(y) = \begin{cases} \frac{1}{2\Gamma(n/2)} \left(\frac{y}{2}\right)^{\frac{n}{2}-1} e^{-\frac{y}{2}}, & y > 0, \\ 0, & y \leq 0, \end{cases}$$

其中, $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$.



χ^2 分布的性质

1. 设 $\chi^2 \sim \chi^2(n)$, 则有 $E(\chi^2) = n$, $\text{Var}(\chi^2) = 2n$;

2. 设 $Y_i \sim \chi^2(n_i)$, $i = 1, 2$, 且 Y_1, Y_2 相互独立, 则有

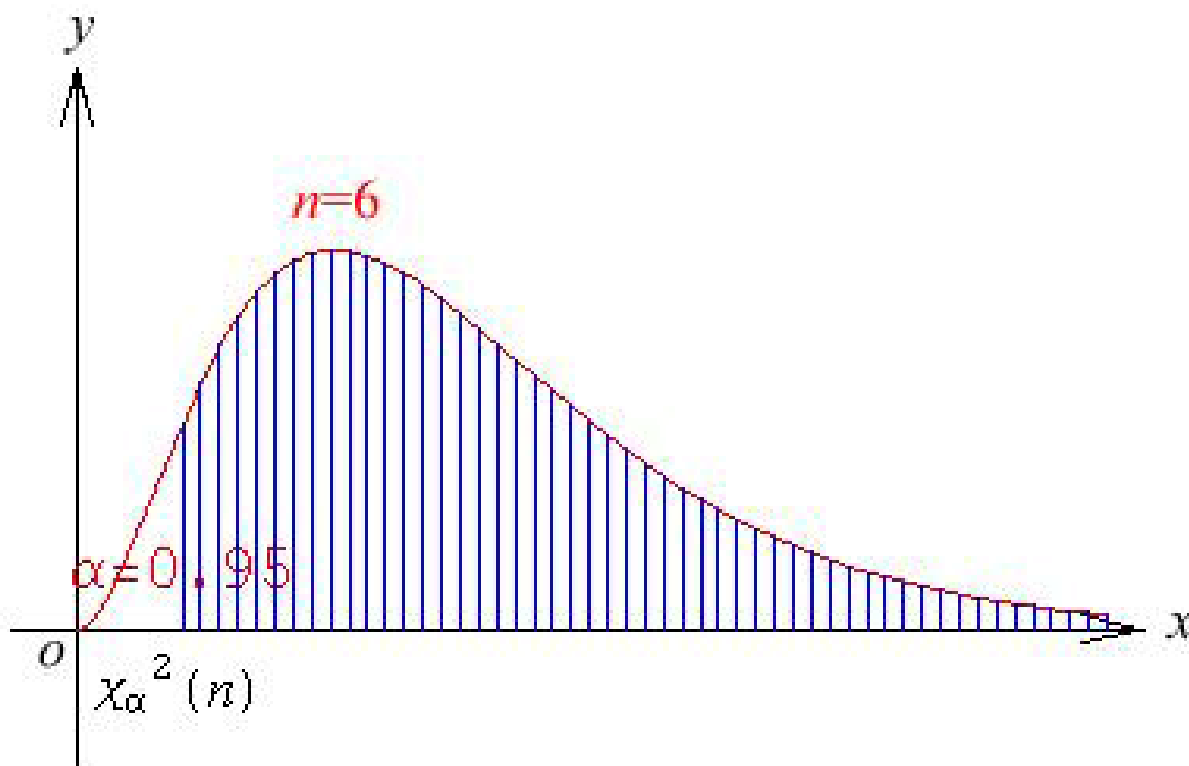
$Y_1 + Y_2 \sim \chi^2(n_1 + n_2)$; —— χ^2 分布可加性

一般地, 若 $Y_i \sim \chi^2(n_i)$, $i = 1, 2, \dots, m$, Y_1, Y_2, \dots, Y_m

相互独立, 则 $\sum_{i=1}^m Y_i \sim \chi^2\left(\sum_{i=1}^m n_i\right)$.



对给定的概率 $\alpha, 0 < \alpha < 1$, 称满足条件 $\int_{\chi_{\alpha}^2(n)}^{\infty} f_n(y) dy = \alpha$ 的点 $\chi_{\alpha}^2(n)$ 为 $\chi^2(n)$ 分布的上 α 分位数, 上 α 分位数 $\chi_{\alpha}^2(n)$ 的值可查 χ^2 分布表



例2.1 利用EXCEL求 $\chi^2_{0.1}(25)$.

在**EXCEL**表单的任一单元格输入

“=CHISQ.INV.RT (0.1,25)” ；

点击"确定" 即在单元格中出现 **"34.382"**.

例2.2 设总体 $X \sim N(\mu, \sigma^2)$, μ, σ^2 已知,

(X_1, X_2, \dots, X_n) 是取自总体 X 的样本.

求(1) 统计量 $\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$ 的分布;

(2) 设 $n = 5$, 若 $a(X_1 - X_2)^2 + b(2X_3 - X_4 - X_5)^2 \sim \chi^2(k)$,
则 a, b, k 各为多少?

解：(1) 作变换 $Y_i = \frac{X_i - \mu}{\sigma} \quad i = 1, 2, \dots, n,$

显然 Y_1, Y_2, \dots, Y_n 相互独立, 且 $Y_i \sim N(0, 1) \quad i = 1, 2, \dots, n,$

$$\text{于是 } \chi^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n Y_i^2 \sim \chi^2(n)$$

$$(2) \quad X_1 - X_2 \sim N(0, 2\sigma^2), \quad \frac{X_1 - X_2}{\sqrt{2}\sigma} \sim N(0, 1)$$

$$2X_3 - X_4 - X_5 \sim N(0, 6\sigma^2), \quad \frac{2X_3 - X_4 - X_5}{\sqrt{6}\sigma} \sim N(0, 1)$$

$$\frac{X_1 - X_2}{\sqrt{2}\sigma} \text{ 与 } \frac{2X_3 - X_4 - X_5}{\sqrt{6}\sigma} \text{ 相互独立,}$$

$$\text{故 } \frac{(X_1 - X_2)^2}{2\sigma^2} + \frac{(2X_3 - X_4 - X_5)^2}{6\sigma^2} \sim \chi^2(2)$$

$$a = \frac{1}{2\sigma^2},$$

$$b = \frac{1}{6\sigma^2},$$

$$k = 2.$$

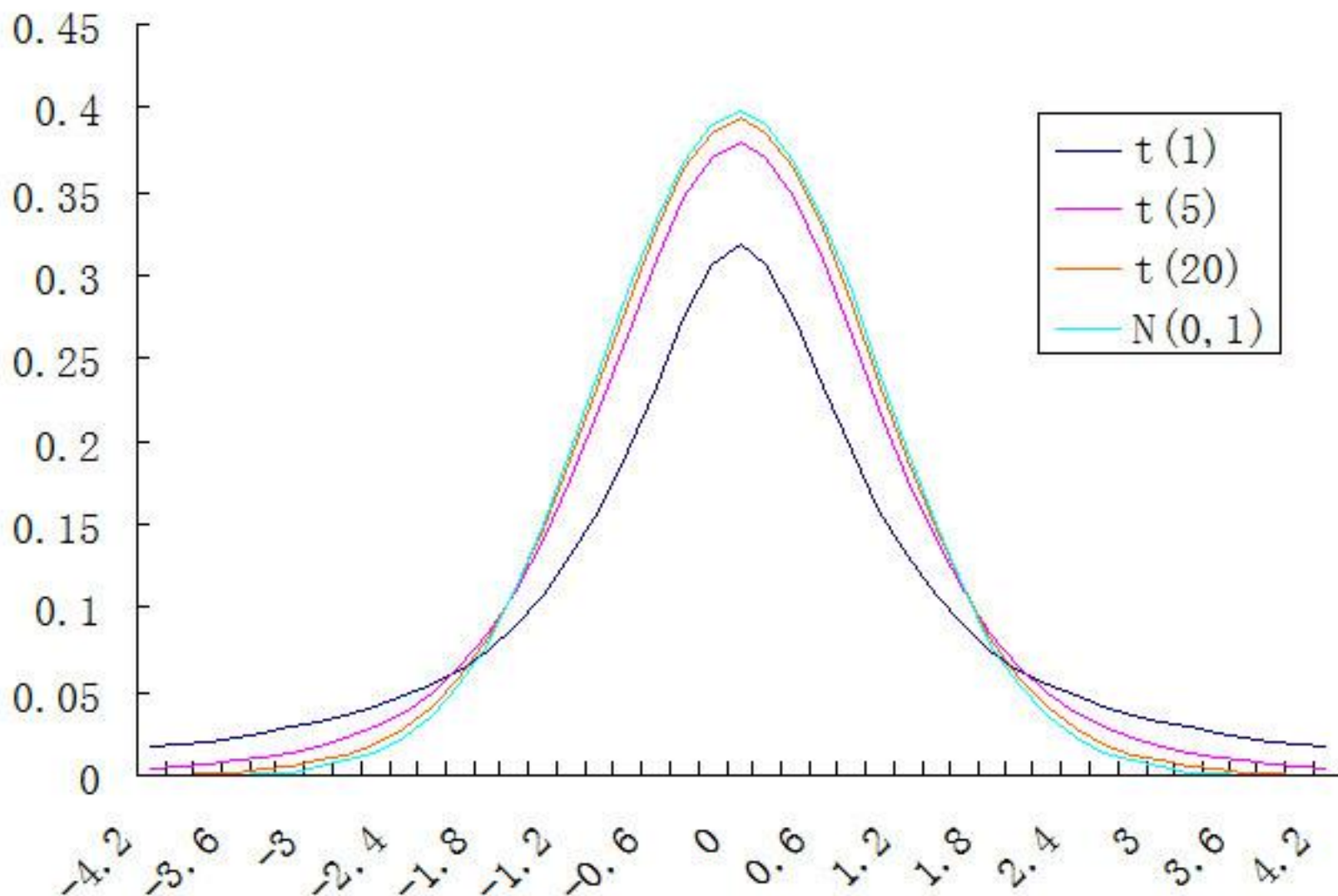
(二) t 分布

设 $X \sim N(0,1)$, $Y \sim \chi^2(n)$, 并且假设 X, Y 相互独立,

则称 $T = \frac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t 分布. 记为 $T \sim t(n)$

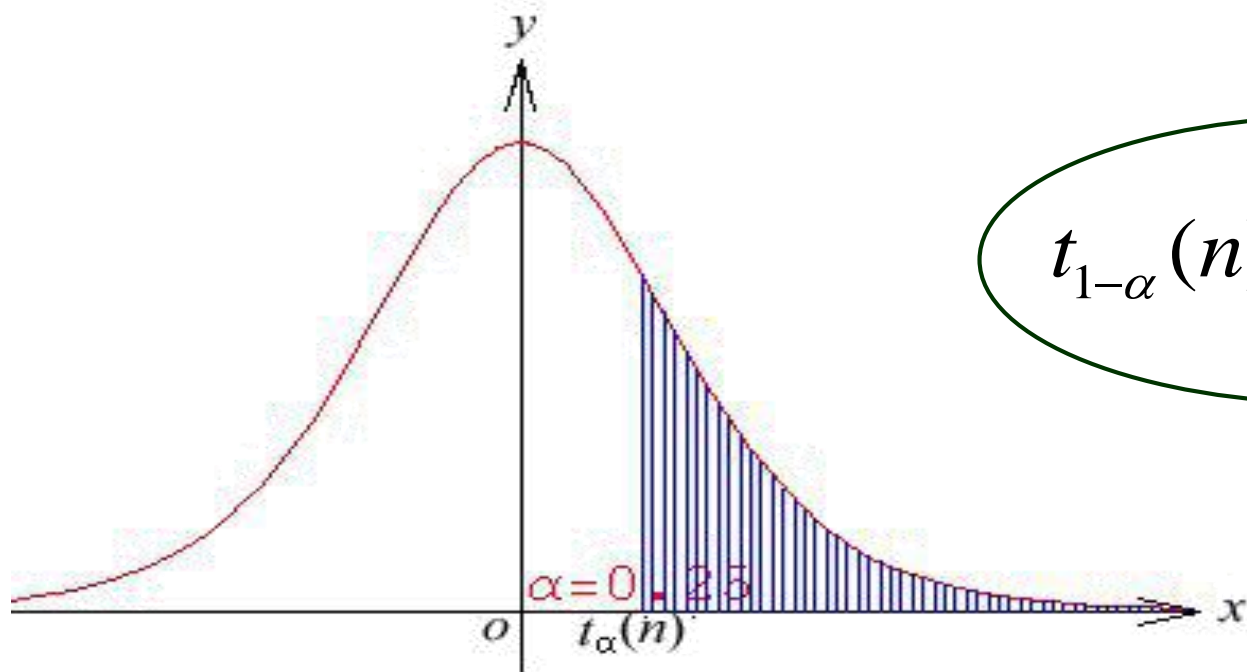
$t(n)$ 分布的概率密度函数为:

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



$t(n)$ 分布概率密度函数

对给定的 α , $0 < \alpha < 1$, 称满足条件 $\int_{t_\alpha(n)}^{\infty} f(t, n) dt = \alpha$ 的点 $t_\alpha(n)$ 为 $t(n)$ 分布的上 α 分位数。 t 分布的上 α 分位数可查 t 分布表



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

例2.3 利用EXCEL求 $t_{0.05}(25)$.

在EXCEL表单的任一单元格输入

**“=T.INV (1-0.05, 25)” 或 “=T.INV.2T
(0.05*2, 25)” ;**

点击“确定” 即在单元格中出现 **“1.708”** .

(三) F 分布

设 $X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$, 且 X, Y 独立, 则

称随机变量 $F = \frac{X/n_1}{Y/n_2}$ 服从自由度 (n_1, n_2) 的 F 分布,

记为 $F \sim F(n_1, n_2)$.

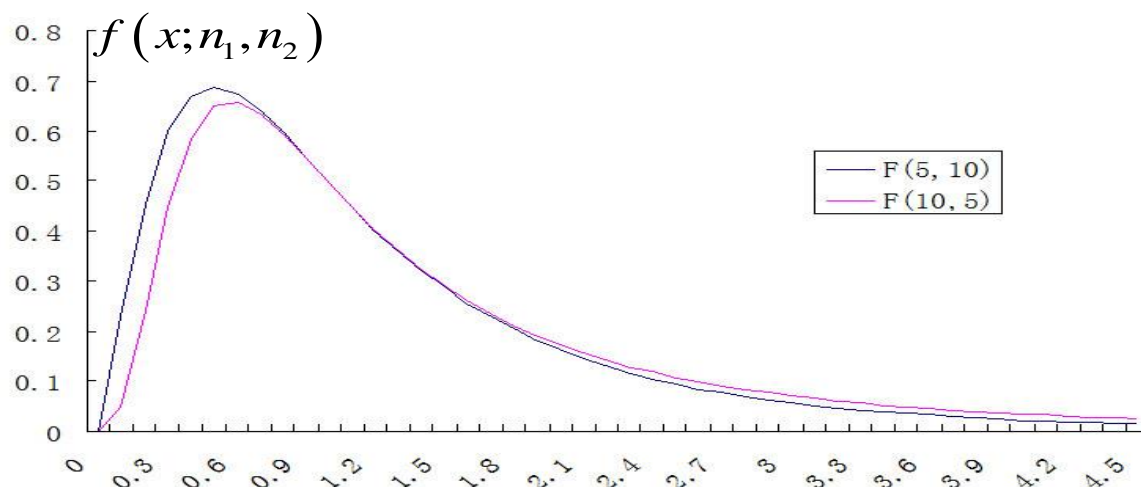
其中, n_1 称为第一自由度, n_2 称为第二自由度.

性质: $F \sim F(n_1, n_2)$, 则 $F^{-1} \sim F(n_2, n_1)$.

$F(n_1, n_2)$ 分布的概率密度函数为:

$$f(x; n_1, n_2) = \begin{cases} \frac{1}{B(n_1/2, n_2/2)} n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}} x^{\frac{n_1}{2}-1} (n_2 + n_1 x)^{-\frac{n_1+n_2}{2}}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

$$\text{其中 } B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$



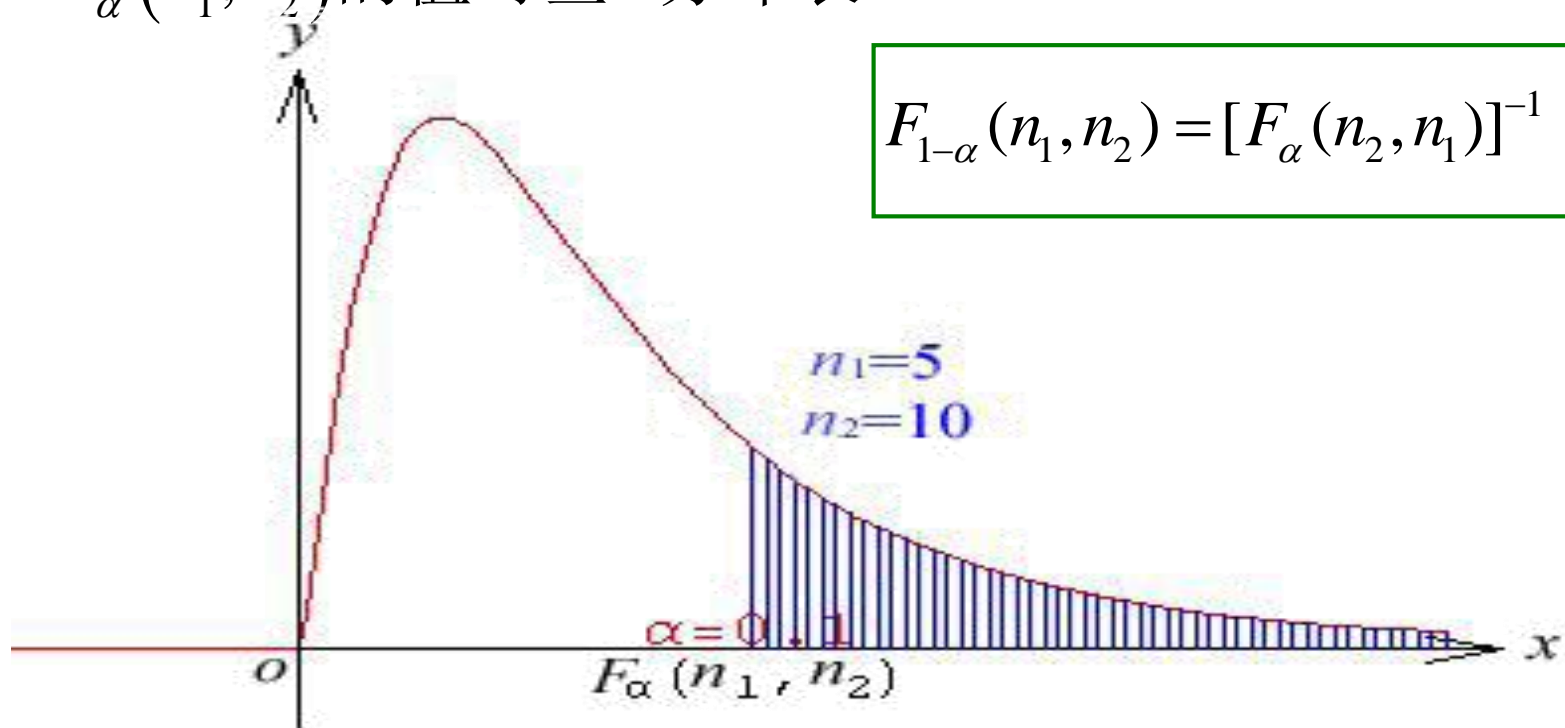
$F(n_1, n_2)$ 分布概率密度函数

对于给定的 α , $0 < \alpha < 1$, 称满足条件

$$\int_{F_\alpha(n_1, n_2)}^{\infty} f(x; n_1, n_2) dx = \alpha$$

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

$F_\alpha(n_1, n_2)$ 的值可查 F 分布表.



例2.4 利用EXCEL求 $F_{0.1}(9,10)$.

在**EXCEL**表单的任一单元格输入

**“=F.INV.RT (0.1, 9, 10)” 或 “=F.INV (1-
0.1, 9, 10)” ;**

点击“确定” 即在单元格中出现 **“2.347” .**

6.3 正态总体下的抽样分布

定理 6.3.1 设 X_1, X_2, \dots, X_n 为来自正态总体

$N(\mu, \sigma^2)$ 的简单随机样本, \bar{X} 是样本均值,

S^2 是样本方差, 则有:

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

定理 6.3.2 设 X_1, X_2, \dots, X_n 为来自正态总体

$N(\mu, \sigma^2)$ 的简单随机样本, \bar{X} 是样本均值,

S^2 是样本方差, 则有:

$$(1) \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

(2) \bar{X} 与 S^2 相互独立.

[思考题]:

设 X_1, X_2, \dots, X_n 是来自正态总体 $N(\mu, \sigma^2)$ 的简单随机样本, \bar{X} 和 S^2 分别是样本均值和样本方差。问:

$$(1) \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \text{ 服从什么分布?}$$

$$(2) \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \text{ 服从什么分布?}$$

答: (1) $\chi^2(n-1)$, (2) $\chi^2(n)$.

定理 6.3.3 设 X_1, X_2, \dots, X_n 为来自正态总体

$N(\mu, \sigma^2)$ 的简单随机样本, \bar{X} 是样本均值, S^2

是样本方差, 则有: $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$.

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

定理6.3.4 设样本 (X_1, \dots, X_{n_1}) 和 (Y_1, \dots, Y_{n_2}) 分别来自总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$, 且相互独立, 样本均值分别为 \bar{X}, \bar{Y} , 样本方差分别为 S_1^2, S_2^2 ,

$$\text{则(1)} \quad \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{S_1^2}{S_2^2} \bigg/ \frac{\sigma_1^2}{\sigma_2^2} \sim F(n_1 - 1, n_2 - 1);$$

$$\text{注意到} \quad \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{(n_1 - 1)S_1^2}{\sigma_1^2(n_1 - 1)} \bigg/ \frac{(n_2 - 1)S_2^2}{\sigma_2^2(n_2 - 1)} \sim F(n_1 - 1, n_2 - 1);$$

$$(2) \quad \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1);$$

(3) 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

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

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \bigg/ \frac{\sqrt{\left(\frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}\right)}}{(n_1 + n_2 - 2)} \sim t(n_1 + n_2 - 2).$$

例3.1 设总体 X 的均值 μ , 方差 σ^2 存在,
 (X_1, \dots, X_n) 是取自总体 X 的样本,
 \bar{X}, S^2 为样本均值和样本方差;
求: (1) $E(\bar{X}), Var(\bar{X}), E(S^2)$;
(2) 若 $X \sim N(\mu, \sigma^2)$, 求 $Var(S^2)$.

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

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

$$E(S^2) = E\left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right) = E\left(\frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right)\right)$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2)\right)$$

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

$$(2) X \sim N(\mu, \sigma^2), \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

$$\Rightarrow \text{Var}\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1)$$

$$\Rightarrow \text{Var}(S^2) = \frac{2\sigma^4}{n-1}.$$

例3.2 设总体 $X \sim N(\mu, \sigma^2)$, (X_1, \dots, X_4)

与 (Y_1, \dots, Y_9) 是取自总体 X 的两个独立样本,

\bar{X}, S_1^2 和 \bar{Y}, S_2^2 分别为样本均值和样本方差;

求 (1) 若 $a \frac{\bar{X} - \bar{Y}}{S_1} \sim t(k)$, 则 a, k 各为多少?

(2) $\sum_{i=1}^4 (X_i - \mu)^2 / 4S_2^2$ 服从什么分布?

(1) $\bar{X} \sim N(\mu, \frac{\sigma^2}{4})$, $\bar{Y} \sim N(\mu, \frac{\sigma^2}{9})$, 且 \bar{X} 与 \bar{Y} 相互独立,

$$\therefore \bar{X} - \bar{Y} \sim N(0, \frac{13\sigma^2}{36}), \quad \frac{6(\bar{X} - \bar{Y})}{\sqrt{13}\sigma} \sim N(0, 1)$$

又 $\frac{3S_1^2}{\sigma^2} \sim \chi^2(3)$, 且 $\bar{X} - \bar{Y}$ 与 S_1^2 相互独立,

由 t 分布定义, $\frac{6}{\sqrt{13}} \frac{\bar{X} - \bar{Y}}{\sigma} \bigg/ \sqrt{\frac{3S_1^2}{3\sigma^2}} = \frac{6\sqrt{13}}{13} \frac{\bar{X} - \bar{Y}}{S_1} \sim t(3)$

$$\Rightarrow a = \frac{6\sqrt{13}}{13}, k = 3.$$

$$(2) \frac{1}{\sigma^2} \sum_{i=1}^4 (X_i - \mu)^2 \sim \chi^2(4), \quad \frac{8S_2^2}{\sigma^2} \sim \chi^2(8),$$

$$\text{且} \sum_{i=1}^4 (X_i - \mu)^2 \text{与} S_2^2 \text{独立},$$

由 F 分布定义知,

$$\frac{1}{4\sigma^2} \sum_{i=1}^4 (X_i - \mu)^2 \bigg/ \frac{8S_2^2}{8\sigma^2} = \sum_{i=1}^4 (X_i - \mu)^2 \bigg/ 4S_2^2 \sim F(4, 8).$$