概率论与数理统计

第六章·样本及抽样分布

2021 年 2 月 24 日

■ 统计与数学学院　■ 王官杰

概率论与数理统计

概　率　论：给定概率分布，研究数据出现概率.

数理统计：给定部分观测数据，研究概率分布.

第一节	总体与样本
第二节	样本分布函数　直方图
第三节	样本函数与统计量
第四节	抽样分布

总体、个体与样本

数理统计中，称研究问题所涉及对象的全体为总体，总体中的每个成员为个体。从总体中抽出的若干个体称为样本。

总体、个体与样本

数理统计中，称研究问题所涉及对象的全体为总体，总体中的每个成员为个体。从总体中抽出的若干个体称为样本。

例 1　研究某工厂生产的电视机的寿命：

- 总体：工厂生产的电视机的全体
- 个体：工厂生产的每台电视机
- 样本：从全部电视机中抽取的一些样品

实际处理中，我们真正关心的并不一定是总体或个体本身，而真正关心的是总体或个体的某项数量指标。故也将总体理解为那些研究对象的某项数量指标的全体。

实际处理中，我们真正关心的并不一定是总体或个体本身，而真正关心的是总体或个体的某项数量指标。故也将总体理解为那些研究对象的某项数量指标的全体。

例 2　研究某工厂生产的电视机的寿命：

- 总体：工厂生产的电视机的寿命的全体
- 个体：工厂生产的每台电视机的寿命

实际处理中，我们真正关心的并不一定是总体或个体本身，而真正关心的是总体或个体的某项数量指标。故也将总体理解为那些研究对象的某项数量指标的全体。

例 2　研究某工厂生产的电视机的寿命：

- 总体：工厂生产的电视机的寿命的全体
- 个体：工厂生产的每台电视机的寿命

例 3　研究某地区所有家庭的年收入：

- 总体：所有家庭的年收入的全体
- 个体：每个家庭的年收入

总体分布

对一个总体，如果用 X 表示其数量指标，则我们随机地抽取个体时，X就构成总体上的一个随机变量。

总体分布

对一个总体，如果用 X 表示其数量指标，则我们随机地抽取个体时，X就构成总体上的一个随机变量。

X 的分布称为总体分布。总体的特性是由总体分布来刻画的。因此，常把总体和总体分布视为同义语。

总体分布

如果总体包含的个体数量是有限的，则称该总体为有限总体。否则称该总体为无限总体。

总体分布

如果总体包含的个体数量是有限的，则称该总体为有限总体。否则称该总体为无限总体。

有限总体的分布是离散型的，且分布通常与总体所含个体数量有关系，研究起来比较困难。

总体分布

如果总体包含的个体数量是有限的，则称该总体为有限总体。否则称该总体为无限总体。

有限总体的分布是离散型的，且分布通常与总体所含个体数量有关系，研究起来比较困难。

故总体所含的个体数量很大时，一般近似视之为无限总体。

样本的二重性

假设$X_1, X_2, \cdots, X_n$是从总体X中取出的样本，

1. 在对这些样本进行观测之前，$X_1, \cdots, X_n$是相互独立的随机变量，均服从总体分布；
2. 一旦对样本进行观测，$X_1, \cdots, X_n$即为确定的一组数值。

从而样本兼有随机变量和确定数值两种属性。有时为了区分，也将$X_1, X_2, \cdots, X_n$的观测值记为$x_1, x_2, \cdots, x_n$，称为样本值。

精确定义

定义 4　称随机变量$X_1, X_2, \cdots, X_n$构成一个（简单）随机样本，如果这些随机变量

1. 相互独立；
2. 服从相同的分布。

它们共同服从的分布称为总体分布；样本个数n称为样本容量。

样本分布

假设总体X服从离散型分布

$$P\{X=x\}=p(x)$$

则$X_1,X_2,\cdots,X_n$的联合分布律为

$$P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}$$

$$=p(x_1)p(x_2)\cdots p(x_n).$$

样本分布

假设总体X服从连续型分布且密度函数为

$$f(x)$$

则$X_1, X_2, \cdots, X_n$的联合概率密度为

$$g(x_1, \cdots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

第一节	总体与样本
第二节	样本分布函数　直方图
第三节	样本函数与统计量
第四节	抽样分布

第一节 总体与样本

第二节 样本分布函数 直方图

第三节 样本函数与统计量

第四节 抽样分布

统计量

在实际问题中，总体分布一般是未知的，我们常常事先假定总体分布的类型，再通过取样的方式确定分布中的未知参数。此时这些未知参数常常写成样本的函数。

定义 5　若样本函数$g(X_1, \ldots, X_n)$不含有任何未知参数，则称这类函数为统计量。

统计量

例如：研究某城市居民的收入情况，事先假定该城市居民的年收入X服从正态分布$N(\mu, \sigma^2)$，其中μ 与σ^2都是未知参数。

在抽取样本$X_1, X_2, \cdots, X_n$的情况下，一般用样本平均值

$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$

近似估计μ，该平均值就是一个统计量。

作为对比，以下函数含有问题中的未知参数，因此不是统计量

$$\frac{X_1 + X_2 + \cdots + X_n}{n\sigma},$$

$$\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu.$$

常用统计量

定义 6　对样本$X_1, X_2, \cdots, X_n$，称

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

为样本均值。

定义 7　对样本$X_1, X_2, \cdots, X_n$，称

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

为样本方差；称

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

为样本标准差。

样本方差的性质：

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

样本方差的性质：

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

例 8　已知样本值为$(2, -1, 0, -2, 0)$，求$\overline{X}$和S^2。

解:　$\overline{X} = -\frac{1}{5}$和$S^2 = \frac{11}{5}$.

样本方差的性质：

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

例 8 已知样本值为$(2, -1, 0, -2, 0)$，求$\overline{X}$和S^2。

解： $\overline{X} = -\frac{1}{5}$和$S^2 = \frac{11}{5}$.

练习 1 已知样本值为$(0, 1, 3, -3, -2)$，求$\overline{X}$和S^2。

解： $\overline{X} = -\frac{1}{5}$和$S^2 = \frac{57}{10}$.

定义 9 对样本$X_1, X_2, \cdots, X_n$及正整数k，称

$$A_k := \frac{1}{n}\sum_{i=1}^{n} X_i^k = \frac{X_1^k + X_2^k + \cdots + X_n^k}{n}$$

为 k阶样本原点矩；对$k \geq 2$，称

$$M_k := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k$$

为 k阶样本中心矩。

均值的大样本分布

中心极限定理的常用结论：

大量同分布随机变量的和、平均值近似服从正态分布。

中心极限定理的常用结论:

大量同分布随机变量的和、平均值近似服从正态分布。

定理 10 设$X_1, X_2, \cdots, X_n$是来自均值为μ、方差为σ^2 的总体的简单样本，则当 n 充分大时，近似地有

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

例 11　用机器向瓶中灌装液体洗净剂，规定每瓶装μ毫升。但实际灌装量总有一定的波动。假定灌装量的方差$\sigma^2 = 1$，如果每箱装这样的洗净剂25瓶。求这25瓶洗净剂的平均灌装量与标定值μ相差不超过0.3毫升的概率．如果每箱装50瓶呢？

第一节	总体与样本
第二节	样本分布函数　直方图
第三节	样本函数与统计量
第四节	**抽样分布**

第四节	抽样分布
4.1	三个重要分布
4.2	正态总体统计量的分布

统计学的三大分布

统计量的分布称为抽样分布。

统计学的三大分布

统计量的分布称为抽样分布。

以下三个来自正态分布的抽样分布

$$\chi^2\text{分布，}t\text{分布，}F\text{分布}$$

称为统计学的三大分布。

定义 12 设$X_1, X_2, \cdots, X_n$相互独立，都服从标准正态分布，则

$$Y := \sum_{i=1}^{n} X_i^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

称为服从n个自由度的 χ^2分布，记为$Y \sim \chi^2(n)$。

定义 12 设$X_1, X_2, \cdots, X_n$相互独立，都服从标准正态分布，则

$$Y := \sum_{i=1}^{n} X_i^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

称为服从n个自由度的 χ^2分布，记为$Y \sim \chi^2(n)$。

定理 13 n个自由度的χ^2分布的概率密度函数为：

$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0; \\ 0, & x \le 0. \end{cases}$$

χ^2分布的密度函数

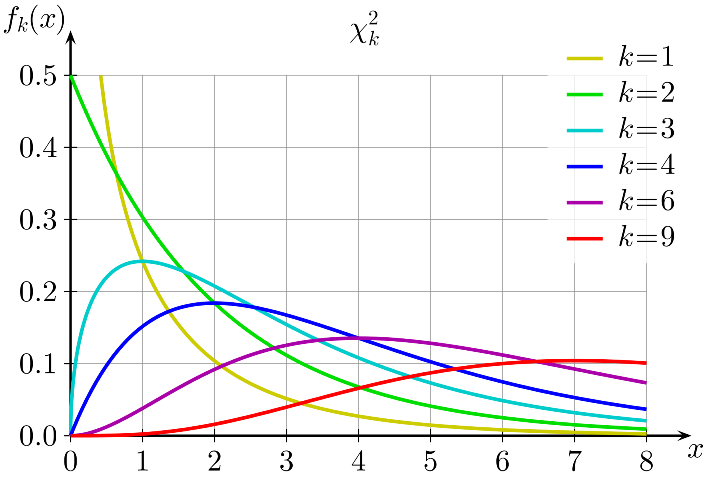


图: χ^2分布的密度函数

χ^2分布的性质：

1. 若X服从标准正态分布，$Y = X^2$，则Y服从1个自由度的χ^2分布，即

$$Y \sim \chi^2(1).$$

χ^2分布

χ^2分布的性质：

1. 若X服从标准正态分布，$Y = X^2$，则Y服从1个自由度的χ^2分布，即
$$Y \sim \chi^2(1).$$
2. 可加性：设$Y_1 \sim \chi^2(m)$，$Y_2 \sim \chi^2(n)$，且两者相互独立，则
$$Y_1 + Y_2 \sim \chi^2(m+n).$$

χ^2分布

χ^2分布的数字特征：

$$E(\chi^2(n)) = n, \quad \mathrm{Var}(\chi^2(n)) = 2n.$$

χ^2分布

定义 14　对给定的$\alpha \in (0, 1)$，称满足条件

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

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

定义 14　对给定的$\alpha \in (0,1)$，称满足条件

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

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

例 15　设$\alpha = 0.05$，$n = 20$，查表得

$$\chi^2_{0.05}(20) = 31.41.$$

定义 16 设两个随机变量X, Y相互独立，并且

$$X \sim N(0,1), \qquad Y \sim \chi^2(n).$$

则称

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

为服从n个自由度的 t分布，记为 $T \sim t(n)$。

定义 16 设两个随机变量X, Y相互独立，并且

$$X \sim N(0,1), \qquad Y \sim \chi^2(n).$$

则称

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

为服从n个自由度的 t分布，记为 $T \sim t(n)$。

定理 17 具有n个自由度的t分布的概率密度函数为：

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

定义 16　设两个随机变量X, Y相互独立，并且

$$X \sim N(0,1), \qquad Y \sim \chi^2(n).$$

则称

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

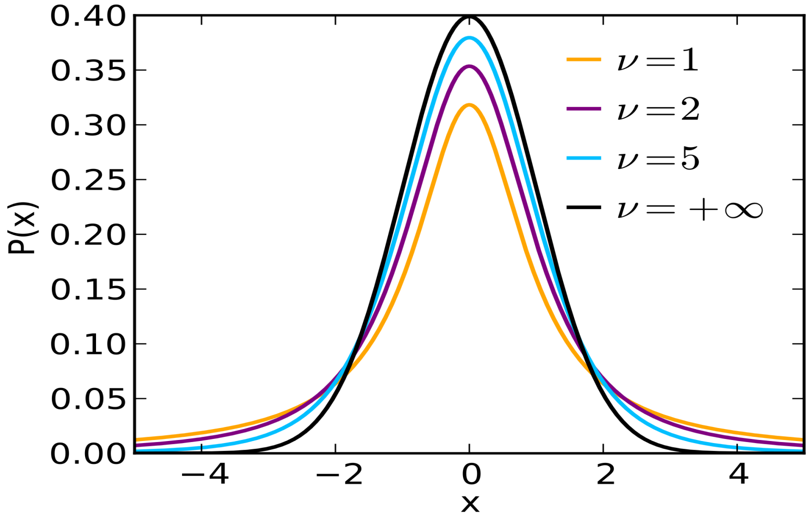
为服从n个自由度的 t分布，记为 $T \sim t(n)$。

定理 17　具有n个自由度的t分布的概率密度函数为：

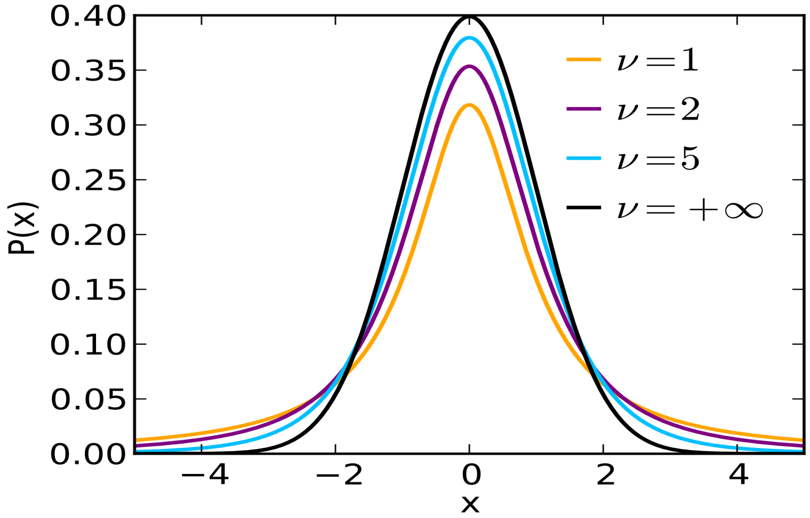
$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

注记　t分布的概率密度函数为偶函数。

t分布的密度函数


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
P(x)
−4
−2
0
2
4
x
$\nu=1$
$\nu=2$
$\nu=5$
$\nu=+\infty$

t分布的密度函数



注记　t分布与标准正态分布的关系：$t(\infty)=N(0,1)$。

t分布

设$T \sim t(n)$。对给定的$\alpha \in (0, 1)$，称满足条件

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

的点$t_\alpha(n)$为 $t(n)$ 分布的上α分位点。

t分布

设$T \sim t(n)$。对给定的$\alpha \in (0,1)$，称满足条件

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

的点$t_\alpha(n)$为 $t(n)$ 分布的上α分位点。

设$Z \sim N(0,1)$，对给定的$\alpha \in (0,1)$，称满足条件

$$P\{Z > Z_\alpha\} = \alpha$$

的点Z_α为标准正态分布的上α分位点。

t分布

设$T \sim t(n)$。对给定的$\alpha \in (0,1)$，称满足条件

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

的点$t_\alpha(n)$为 $t(n)$ 分布的上α分位点。

设$Z \sim N(0,1)$，对给定的$\alpha \in (0,1)$，称满足条件

$$P\{Z > Z_\alpha\} = \alpha$$

的点Z_α为标准正态分布的上α分位点。

例 18　$t_{0.05}(10) = 1.812$，　$Z_{0.025} = 1.960$。

设$T \sim t(n)$。对给定的$\alpha \in (0, 1)$，称满足条件

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

的点$t_\alpha(n)$为 $t(n)$ 分布的上α分位点。

设$Z \sim N(0, 1)$，对给定的$\alpha \in (0, 1)$，称满足条件

$$P\{Z > Z_\alpha\} = \alpha$$

的点Z_α为标准正态分布的上α分位点。

例 18　$t_{0.05}(10) = 1.812$，　$Z_{0.025} = 1.960$。

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

t分布

1876年，Friedrich Helmert 和 Jacob Lüroth 最先推导和证明了t分布.

t分布

1876年，Friedrich Helmert 和 Jacob Lüroth 最先推导和证明了t分布.

1908年，英国人威廉·戈塞(Willam Gosset)再次发现并发表了t分布．当时他在爱尔兰都柏林的啤酒酿酒厂工作．由于酒厂禁止员工发表与酿酒研究有关的成果，他的论文使用了“学生”(Student)作为笔名．因此t分布又称为学生分布。

定义 19　设两个随机变量Y_1, Y_2相互独立，并且

$$Y_1 \sim \chi^2(m), \qquad Y_2 \sim \chi^2(n)$$

则

$$F := \frac{Y_1/m}{Y_2/n} \sim F(m, n).$$

称为自由度为m和n的 F分布，记为 $F \sim F(m, n)$。

定义 19　设两个随机变量Y_1, Y_2相互独立，并且

$$Y_1 \sim \chi^2(m), \qquad Y_2 \sim \chi^2(n)$$

则

$$F := \frac{Y_1/m}{Y_2/n} \sim F(m, n).$$

称为自由度为m和n的 F分布，记为 $F \sim F(m, n)$。

定理 20　自由度为m和n的F分布的概率密度为 $f(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\cdot\Gamma} \\ 0, \end{cases}$

F分布的密度函数

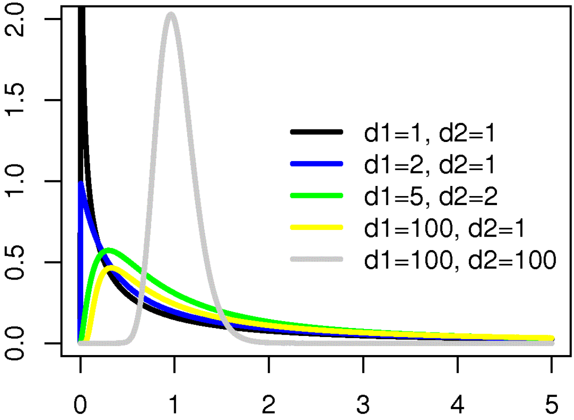


图: F分布的密度函数

F分布的性质

F分布的性质：

1. 若$F \sim F(m, n)$，则 $1/F \sim F(n, m)$。

F分布的性质

F分布的性质：

1. 若$F \sim F(m, n)$，则 $1/F \sim F(n, m)$。
2. 若$T \sim t(n)$，则$T^2 \sim F(1, n)$。

F分布的分位点

设$F \sim F(m, n)$。对给定的$\alpha \in (0, 1)$，称满足条件

$$P\{F > F_\alpha(m, n)\} = \alpha$$

的点$F_\alpha(m, n)$)为$F(m, n)$分布的上α分位点。

F分布的分位点

设$F \sim F(m,n)$。对给定的$\alpha \in (0,1)$，称满足条件

$$P\{F > F_\alpha(m,n)\} = \alpha$$

的点$F_\alpha(m,n)$)为$F(m,n)$分布的上α分位点。

性质 $$F_{1-\alpha}(m,n) = \frac{1}{F_\alpha(n,m)}.$$

F分布的分位点

设$F \sim F(m,n)$。对给定的$\alpha \in (0,1)$，称满足条件

$$P\{F > F_\alpha(m,n)\} = \alpha$$

的点$F_\alpha(m,n)$)为$F(m,n)$分布的上α分位点。

性质 $F_{1-\alpha}(m,n) = \dfrac{1}{F_\alpha(n,m)}$.

例 21 $F_{0.95}(15,10)$

F分布的分位点

设$F \sim F(m,n)$。对给定的$\alpha \in (0,1)$，称满足条件

$$P\{F > F_\alpha(m,n)\} = \alpha$$

的点$F_\alpha(m,n)$)为$F(m,n)$分布的上α分位点。

性质　$F_{1-\alpha}(m,n) = \dfrac{1}{F_\alpha(n,m)}.$

例 21　$F_{0.95}(15,10) = 1/F_{0.05}(10,15)$

F分布的分位点

设$F \sim F(m, n)$。对给定的$\alpha \in (0, 1)$，称满足条件

$$P\{F > F_\alpha(m, n)\} = \alpha$$

的点$F_\alpha(m, n)$)为$F(m, n)$分布的上α分位点。

性质 $$F_{1-\alpha}(m, n) = \frac{1}{F_\alpha(n, m)}.$$

例 21 $F_{0.95}(15, 10) = 1/F_{0.05}(10, 15) = 1/2.54 = 0.394.$

第四节	抽样分布
4.1	三个重要分布
4.2	正态总体统计量的分布

定理22 设$X_1, X_2, \cdots, X_n$是取自正态总体 $N(\mu, \sigma^2)$ 的样本。则$\overline{X}$与S^2相互独立，且有

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1), \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

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

定理 23 设$X_1, X_2, \cdots, X_m$与$Y_1, Y_2, \cdots, Y_n$分别是取自两个相互独立的正态总体

$$N(\mu_1, \sigma^2), \qquad N(\mu_2, \sigma^2)$$

的样本。则

$$U := \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1),$$

其中$\overline{X}, \overline{Y}$ 分别是两个样本各自的均值。

定理 24　设$X_1, X_2, \cdots, X_m$与$Y_1, Y_2, \cdots, Y_n$分别是取自两个相互独立的正态总体

$$N(\mu_1, \sigma^2), \qquad N(\mu_2, \sigma^2)$$

的样本。则

$$T := \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}} \cdot \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2),$$

其中$\overline{X}$, $\overline{Y}$, S_1^2, S_2^2分别是两个样本各自的均值及方差。

两个正态总体的统计量的分布

定理 25　设$X_1, \cdots, X_m$与$Y_1, \cdots, Y_n$分别是取自两个相互独立的正态总体

$$N(\mu_1, \sigma_1^2), \qquad N(\mu_2, \sigma_2^2)$$

的样本。则

$$F := \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(m-1, n-1).$$