概率论与数理统计

第六章 · 样本与统计量

2020 年 2 月 12 日

暨南大学数学系 ■ 吕荐瑞

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	统计量
第三节	统计中的常用分布
第四节	正态统计量的分布

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

在实际研究中，我们真正关注的

- 并不一定是总体或个体本身，
- 而是总体或个体的某项数量指标.

故也将总体理解为研究对象的某项数量指标的全体.

在实际研究中，我们真正关注的

- 并不一定是总体或个体本身，
- 而是总体或个体的某项数量指标.

故也将总体理解为研究对象的某项数量指标的全体.

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寿命的全体
- 个体：工厂生产的每台电视机的寿命

在实际研究中，我们真正关注的

- 并不一定是总体或个体本身，
- 而是总体或个体的某项数量指标.

故也将总体理解为研究对象的某项数量指标的全体.

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寿命的全体
- 个体：工厂生产的每台电视机的寿命

例 2　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词.

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 1　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	**统计量**
第三节	统计中的常用分布
第四节	正态统计量的分布

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 1　样本的已知函数（不含问题中的未知参数）称为统计量．

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.

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 2　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.

常用统计量

定义 3　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.

常用统计量

样本方差的性质：

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

常用统计量

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

常用统计量

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

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

常用统计量

定义 4 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 $k \geqslant 2$，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.

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.

定理 1　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

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

均值的大样本分布

例 2　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？

复习与提高

选择　设总体 $X \sim B(1,p)$，其中参数 $p \in (0,1)$ 未知. X_1,X_2,X_3 是来自总体 X 的简单随机样本，$\overline{X}$ 为样本均值，则下列选项中不是统计量的为 . . . . . (　　)

(A) $\min\{X_1,X_2,X_3\}$　　(B) $X_1-(1-p)\overline{X}$

(C) $\max\{X_1,X_2,X_3\}$　　(D) $X_3-3\overline{X}$

第一节 总体与样本

第二节 统计量

第三节 统计中的常用分布

第四节 正态统计量的分布

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

χ^2 分布，t 分布，F 分布

称为统计学的三大分布.

定义 1 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 $Y \sim \chi_n^2$ 或 $Y \sim \chi^2(n)$.

定义 1　设 $X_1, X_2, \cdots, X_n$ 相互独立，都服从标准正态分布，则

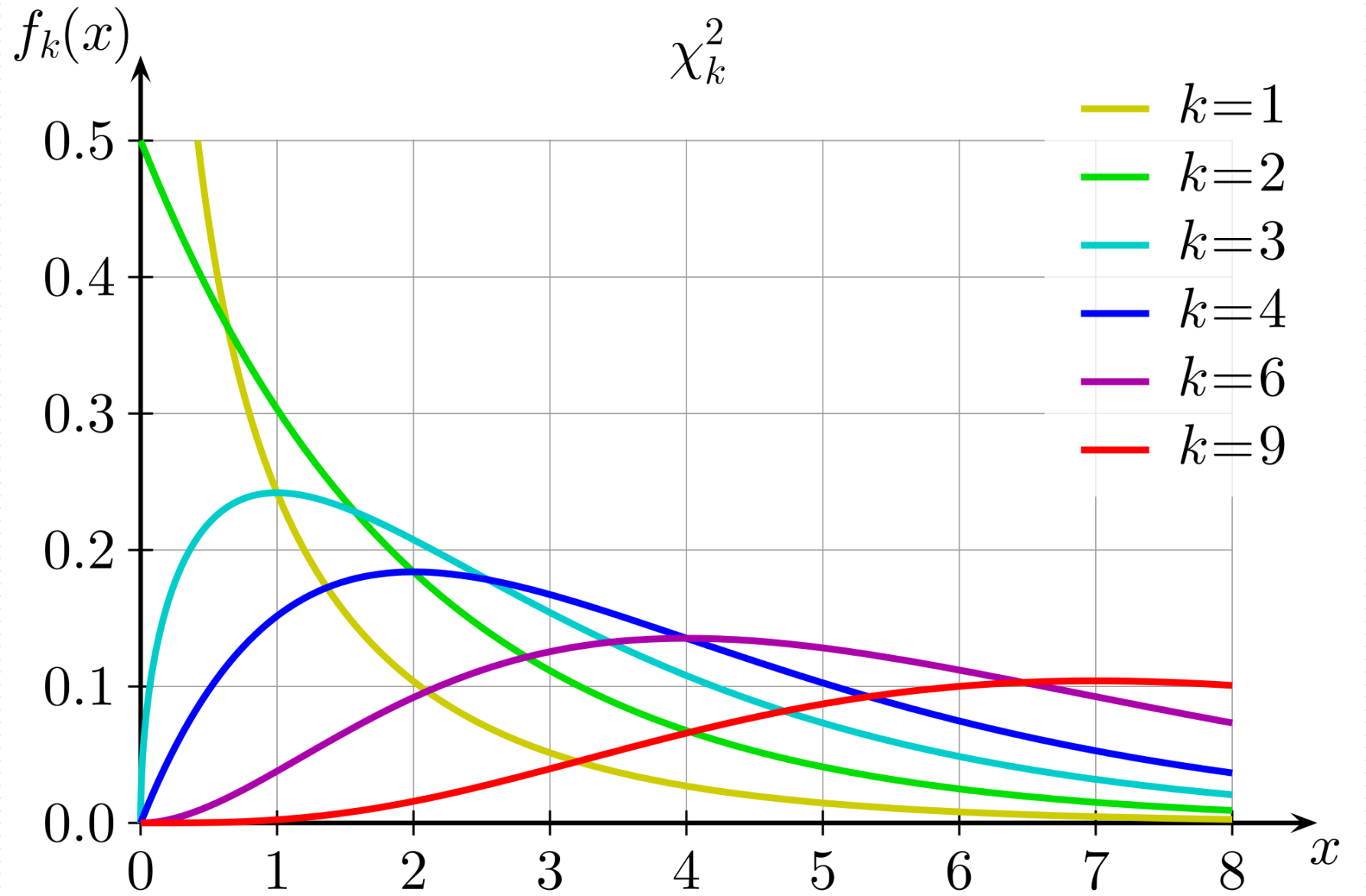
$$Y := \sum_{i=1}^{n} X_i^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

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

定理 1　n 个自由度的 χ^2 分布的概率密度函数为：

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

χ^2 分布的密度函数



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

$$Y \sim \chi_1^2.$$

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

$$Y \sim \chi_1^2.$$

2. 可加性：设 $Y_1 \sim \chi_m^2$，$Y_2 \sim \chi_n^2$，且两者相互独立，则

$$Y_1 + Y_2 \sim \chi_{m+n}^2.$$

χ^2 分布

χ^2 分布的数字特征：

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

χ^2 分布

定义 2 对给定的 $\alpha \in (0,1)$，称满足条件

$$P\{\chi_n^2 > \chi_n^2(\alpha)\} = \alpha$$

的点 $\chi_n^2(\alpha)$ 为 χ_n^2 分布的上 α 分位点.

χ^2 分布

定义 2 对给定的 $\alpha \in (0,1)$，称满足条件

$$P\{\chi_n^2 > \chi_n^2(\alpha)\} = \alpha$$

的点 $\chi_n^2(\alpha)$ 为 χ_n^2 分布的上 α 分位点.

例 1 设 $\alpha = 0.05$，$n = 20$，查表得

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

定义 3 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 $T \sim t_n$ 或 $T \sim t(n)$.

定义 3　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 $T \sim t_n$ 或 $T \sim t(n)$.

定理 2　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 3　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

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

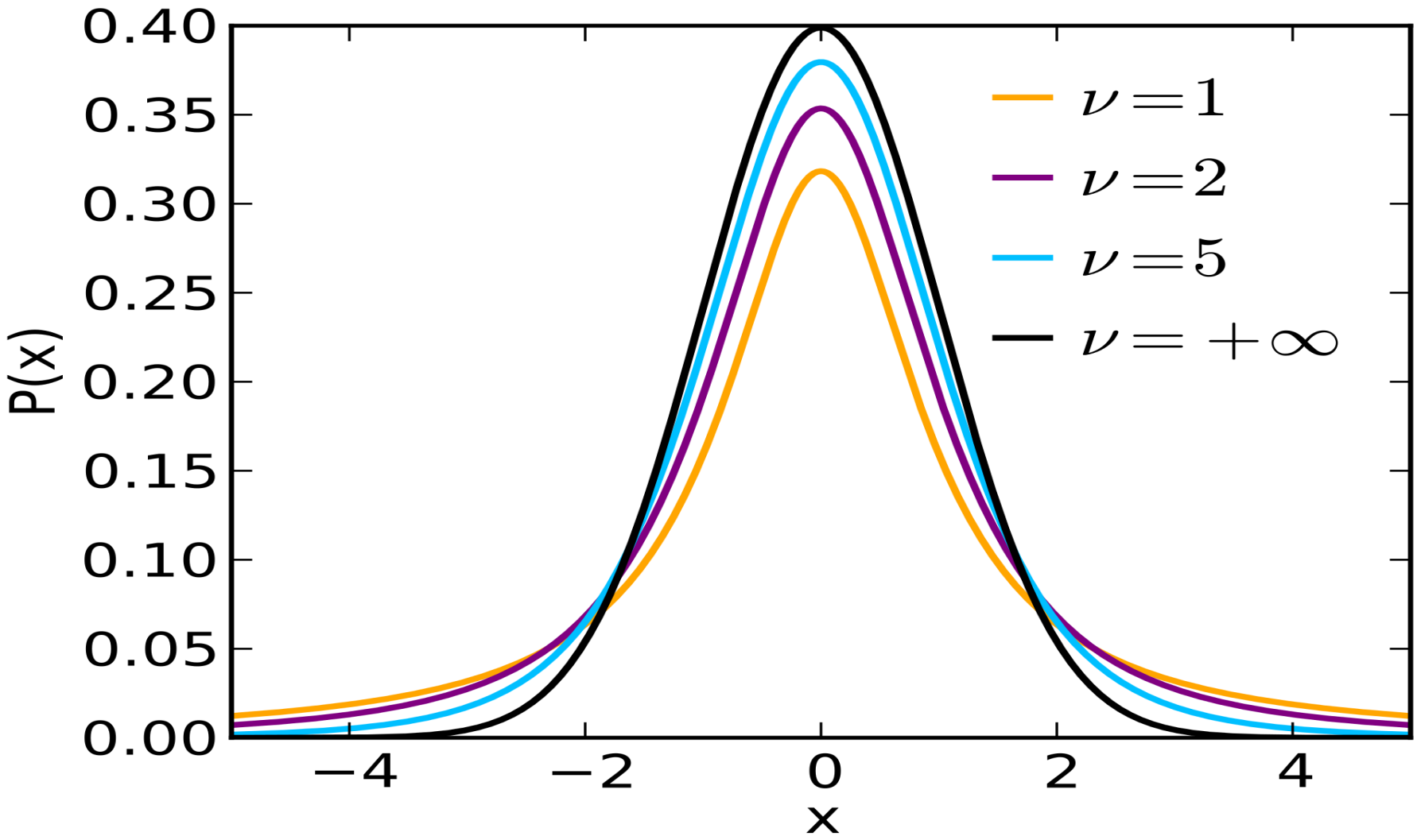
为服从 n 个自由度的 t 分布，记为 $T \sim t_n$ 或 $T \sim t(n)$.

定理 2　具有 n 个自由度的 t 分布的概率密度函数为：

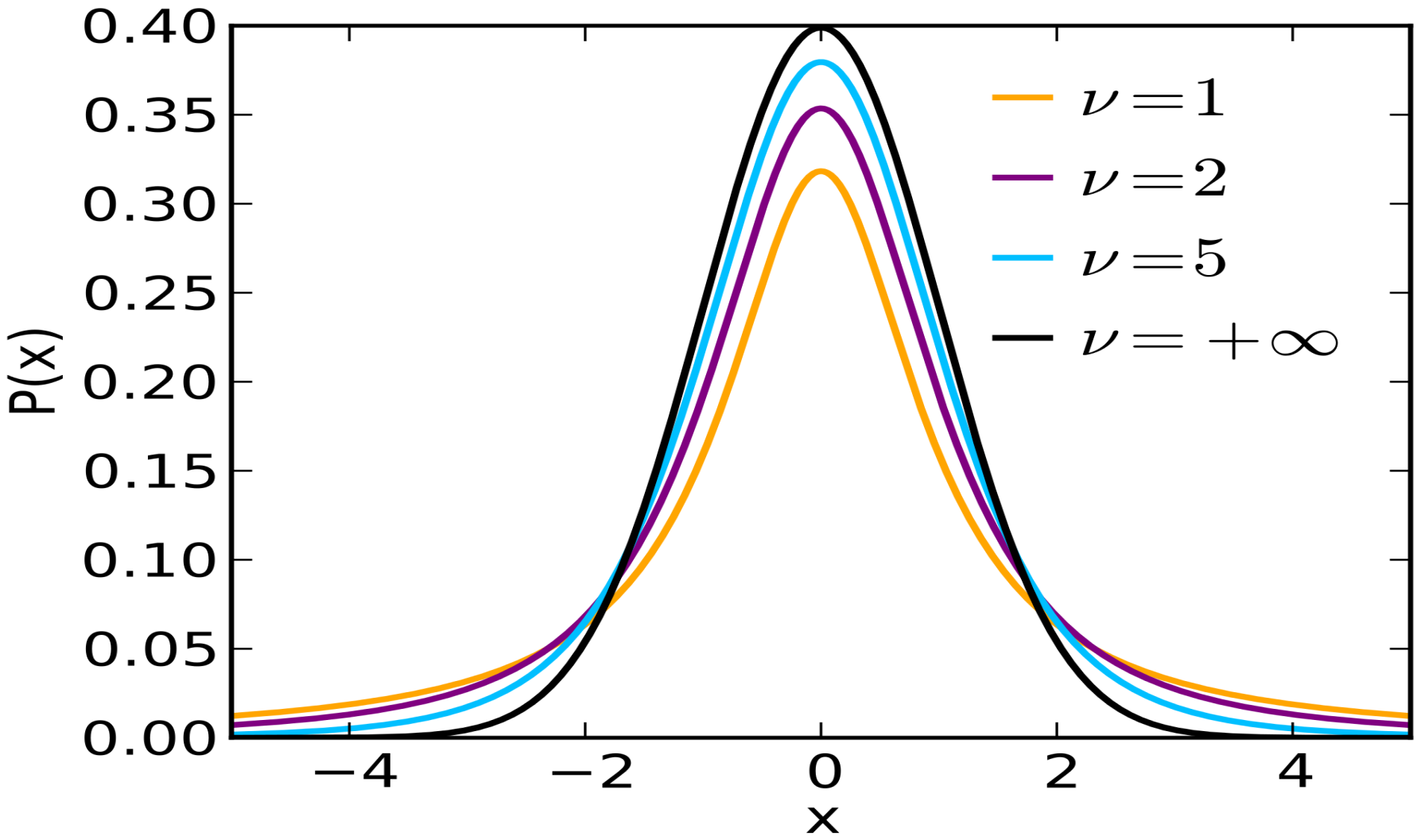
$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

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

$$P\{T > t_n(\alpha)\} = \alpha$$

的点 $t_n(\alpha)$ 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

$$P\{T > t_n(\alpha)\} = \alpha$$

的点 $t_n(\alpha)$ 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

$$P\{T > t_n(\alpha)\} = \alpha$$

的点 $t_n(\alpha)$ 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 2　$t_{10}(0.05) = 1.812$，　$Z_{0.025} = 1.960$.

t 分布

设 $T \sim t_n$. 对给定的 $\alpha \in (0,1)$，称满足条件

$$P\{T > t_n(\alpha)\} = \alpha$$

的点 $t_n(\alpha)$ 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 2　$t_{10}(0.05) = 1.812$，　$Z_{0.025} = 1.960$.

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

定义 4 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 $F \sim F_{m,n}$ 或 $F \sim F(m,n)$.

定义 4 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

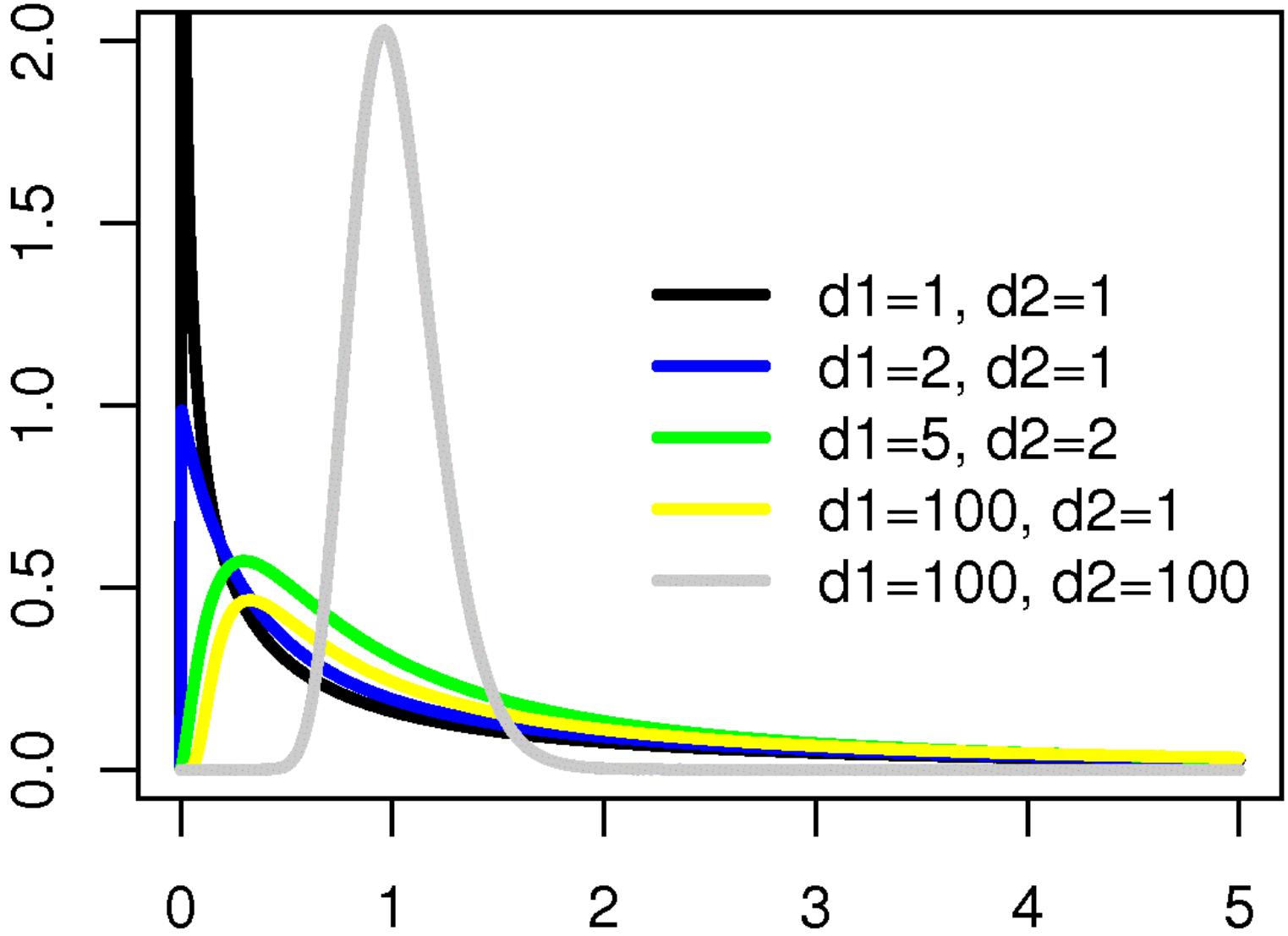
$$F := \frac{Y_1/m}{Y_2/n} \sim F_{m,n}.$$

称为自由度为 m 和 n 的 F 分布，记为 $F \sim F_{m,n}$ 或 $F \sim F(m,n)$.

定理 3 自由度为 m 和 n 的 F 分布的概率密度为

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

F 分布的密度函数


0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
d1=1, d2=1
d1=2, d2=1
d1=5, d2=2
d1=100, d2=1
d1=100, d2=100

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

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

的点 $F_{m,n}(\alpha)$ 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

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

的点 $F_{m,n}(\alpha)$ 为 $F_{m,n}$ 分布的上 α 分位点.

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

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

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

的点 $F_{m,n}(\alpha)$ 为 $F_{m,n}$ 分布的上 α 分位点.

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

例 3 $F_{15,10}(0.95)$

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

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

的点 $F_{m,n}(\alpha)$ 为 $F_{m,n}$ 分布的上 α 分位点.

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

例 3 $F_{15,10}(0.95) = 1/F_{10,15}(0.05)$

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

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

的点 $F_{m,n}(\alpha)$ 为 $F_{m,n}$ 分布的上 α 分位点.

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

例 3 $F_{15,10}(0.95) = 1/F_{10,15}(0.05) = 1/2.54 = 0.394$.

复习与提高

选择　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 $R = \dfrac{X_1 - X_2}{\sqrt{2}|X_3|}$ 服从分布 . . . (　　)

(A) $F_{1,1}$　　(B) $F_{2,1}$　　(C) t_1　　(D) t_2

第一节	总体与样本
第二节	统计量
第三节	统计中的常用分布
第四节	正态统计量的分布

单个正态总体的统计量的分布

定理 1 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

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

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1},$$

$$\frac{\sum_{i=1}^{n}\left(X_i-\mu\right)^2}{\sigma^2} \sim \chi_n^2,$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

两个正态总体的统计量的分布

定理 2 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

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

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.

两个正态总体的统计量的分布

定理 3 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 5 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

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

其中 S_1^2 和 S_2^2 分别是两个样本各自的方差.