第六章　样本及抽样分布

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.

6.1　总体与样本

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. 从总体中抽取若干个个体的过程称为抽样, 从总体中抽出的若干个体称为样本, 样本中所含个体的数量称为样本容量.

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.

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

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．

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．

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．

一个抽样方法被称为简单随机抽样，如果该抽样方法所得到的样本具有：

1. 随机性：总体中每一个个体都有同等机会被选入样本, 这意味着每一样品 X_i 与总体 X 同分布.
2. 独立性：样本中每一个样品取值不影响其他样品的取值, 也不受其他样品取值的影响, 这意味着 $X_1, X_2, \cdots, X_n$ 相互独立.

由简单随机抽样得到的样本称为简单随机样本.

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

$$P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}$$
$$=p(x_1)p(x_2)\cdots p(x_n).$$

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

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

6.2 样本分布函数 直方图

我们把总体的分布函数

$$F(x)=P(X\le x)$$

称为总体分布函数. 从总体中抽取容量为 n 的样本得到 n 个样本观测值, 若样本容量 n 较大，则相同的 n 观测值可能重复出现若干次，为此.

将观测值整理, 并写出下面的样本频率分布表:

观测值	$x_{(1)}$	$x_{(2)}$	$\cdots$	$x_{(l)}$	总计
频数	n_1	n_2	$\cdots$	n_l	n
频率	f_1	f_2	$\cdots$	f_l	1

其中 $x_{(1)}<x_{(2)}<\cdots<x_{(l)}$ $(l\leqslant n)$,

$$f_i=\frac{n_i}{n}\ (i=1,2,\cdots,l),\ \sum_{i=1}^{l}n_i=n,\ \sum_{i=1}^{l}f_i=1.$$

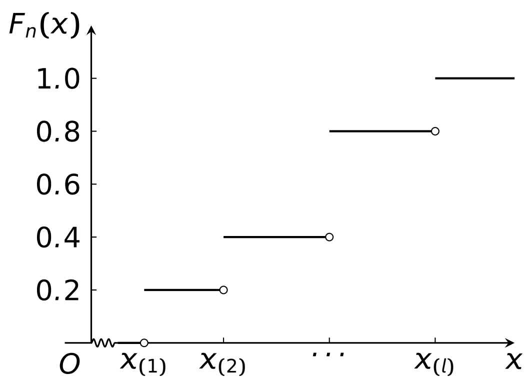
定义 **1.** 设函数

$$F_n(x)=\begin{cases}0, & x<x_{(1)}\\ \sum\limits_{x_{(i)}\leqslant x}f_i, & x_{(i)}\leqslant x<x_{(i+1)}, \quad (i=1,2,\cdots,l-1)\\ 1, & x\geqslant x_{(l)}\end{cases}$$

其中和式 $\sum\limits_{x_{(i)}\leqslant x}$ 是对小于或等于 x 的一切 $x_{(i)}$ 的频率 f_i 求和, 则称 $F_n(x)$ 为样本分布函数或经验分布函数.

样本分布函数 $F_n(x)$ 具有下列性质:

1. $0 \leqslant F_n(x) \leqslant 1$;
2. $F_n(x)$ 是非减函数;
3. $F_n(-\infty) = 0$, $F_n(+\infty) = 1$;
4. $F_n(x)$ 在每个观测值 $x_{(i)}$ 处是右连续的, 点 $x_{(i)}$ 是 $F_n(x)$ 的跳跃间断点, $F_n(x)$ 在该点的跃度就等于频率 f_i,



样本分布函数

对于任意的实数 x, 总体分布函数 $F(x)$ 是事件 $\{X \leqslant x\}$ 的概率; 样本分布函数 $F_n(x)$ 是事件 $\{X \leqslant x\}$ 的频率. 根据伯努利大数定律可知, 当 $n \to \infty$ 时, 对于任意的正数 ε, 有

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

定理. 格利文科定理 设 $x_1, x_2, \ldots, x_n$ 是取自总体分布函数为 $F(x)$ 的样本, $F_n(x)$ 是其经验分布函数, 当 $n \to \infty$ 时, 有

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

该定理表明, 当 n 相当大时, 样本分布函数是总体分布函数 $F(x)$ 的一个良好的近似.

作频率分布直方图的步骤:

1. 找出样本观测值 $x_1, x_2, \cdots, x_n$ 中的最小值与最大值, 分别记作 x_1^* 与与 x_n^*, 即
$$x_1^* = \min\{x_1, x_2, \cdots, x_n\}, \quad x_n^* = \max\{x_1, x_2, \cdots, x_n\}.$$
2. 适当选取略小于 x_1^* 的数 a 与略大于 x_n^* 的数 b, 并用分点
$$a = t_0 < t_1 < t_2 < \cdots < t_{l-1} < t_l = b$$
把区间 (a, b) 分成 l 个子区间
$$[t_0, t_1), [t_1, t_2), \cdots, [t_{i-1}, t_i), \cdots, [t_{l-1}, t_l).$$
第 i 个子区间的长度为 $\Delta t_i = t_i - t_{i-1}, i = 1, 2, \cdots, l$.
3. 把所有样本观测值逐个分到各子区间内, 并计算样本观测值落在各子区间内的频数 n_i 及频率 $f_i = \frac{n_i}{n}(i = 1, 2, \cdots, l)$.
4. 在 Ox 轴上截取各子区间, 并以各子区间为底, 以 $\frac{f_i}{t_i - t_{i-1}}$ 为高作小矩形, 这样作出的所有小矩形就构成了直方图

注记**.** *(1)* 各个小矩形的面积 ΔS_i 就等于样本观测值落在该子区间内的频率, 即
$$\Delta S_i = (t_i - t_{i-1})\frac{f_i}{t_i - t_{i-1}} = f_i \quad (i = 1, 2, \cdots, l).$$

(2) 所有小矩形的面积的和等于 *1:*
$$\sum_{i=1}^{l} \Delta S_i = \sum_{i=1}^{l} f_i = 1.$$

例 **1.** 为研究某厂工人生产某种产品的能力, 我们随机调查了 *20* 位工人某天生产的该种产品的数量, 数据如下

160	*196*	*164*	*148*	*170*
175	*178*	*166*	*181*	*162*
161	*168*	*166*	*162*	*172*
156	*170*	*157*	*162*	*154*

写出产品数量的频率分布表，并作直方图.

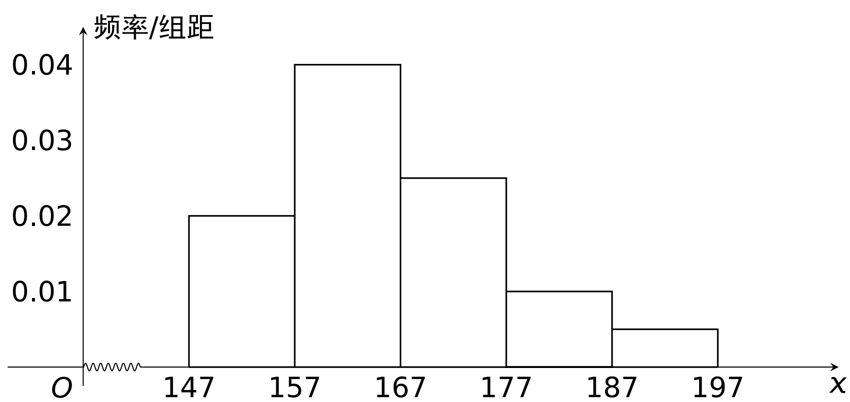
解. 因为样本观测值中最小值为 *148*,最大值为 *196*,所以我们把数据的分布间确定为 $(147,197)$,并将区间分为 *5* 个子区间

$$[147,157),[157,167),[167,177),[177,187),[187,197),$$

由此得频率分布表：

组序	分组区间	频数	频率
1	*[147,157)*	*4*	*0.20*
2	*[157,167)*	*8*	*0.40*
3	*[167,177)*	*5*	*0.25*
4	*[177,187)*	*2*	*0.10*
5	*[187,197)*	*1*	*0.05*
合计		*20*	*1*

根据频率分布表作出直方图：



6.3　样本函数与统计量

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.** 若样本函数 $g(X_1,...,X_n)$ 不含有任何未知参数，则称这类函数为统计量.

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

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

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.

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.

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.

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 样本 k 阶原点矩；对 $k \geq 2$，称

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

为 样本 k 阶中心矩.

大数定律的结论：大量同分布随机变量的算数平均数依概率收敛于它们的期望.

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，总体的 k 阶原点矩存在且为 $E(X^k) = \mu_k$, 则

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mu_k, \quad k = 1,2,\cdots,$$

注记**.** 由第五章中关于依概率收敛的序列的性质知道

$$g\left(A_1,A_2,\cdots,A_k\right) \xrightarrow{P} g\left(\mu_1,\mu_2,\cdots,\mu_k\right),$$

其中 g 为连续函数.

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 **2.** 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

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

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 *(B)*

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

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

6.4 抽样分布

6.4.1 三个重要分布

统计量的分布称为抽样分布.

在使用统计量进行统计推断时常需知道它的分布. 当总体的分布函数已知时，抽样分布是确定的，然而要求出统计量的精确分布，一般来说是困难的.

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

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

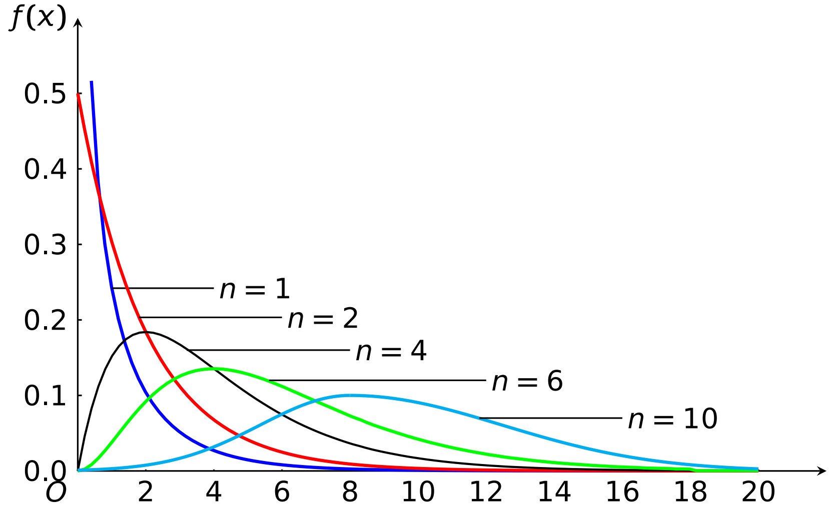
称为服从 n 个自由度的 χ^2 分布，记为 $\chi^2 \sim \chi^2(n)$. 此处的自由度指定义右端包含独立随机变量的个数.

定理 **1.** n 个自由度的 χ^2 分布的概率密度函数为：

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

χ^2 分布的性质：

1. 若 X 服从标准正态分布，$\chi^2 = X^2$，则 χ^2 服从 1 个自由度的 χ^2 分布，即
$$\chi^2 \sim \chi^2(1).$$
2. 可加性：设 $\chi_1^2 \sim \chi^2(n_1)$, $\chi_2^2 \sim \chi^2(n_2)$，且两者相互独立，则
$$\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2).$$
注记**.** 此结论可推广: 设 $X_i \sim \chi^2(n_i)\,(i = 1, 2, \cdots, k)$ 且相互独立, 则
$$\sum_{i=1}^{k} X_i \sim \chi^2\left(\sum_{i=1}^{k} n_i\right).$$



χ^2 分布的数字特征:

$$E(\chi^2(n)) = n,\ D(\chi^2(n)) = 2n.$$

证明. 因 $X_i \sim N(0,1)$, 故 $E\left(X_i^2\right) = D\left(X_i\right) = 1,\ E\left(X_i^4\right) = 3,\ i = 1, 2, \cdots, n$, 因此

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

又

$$D\left(X_i^2\right) = E\left(X_i^4\right) - \left[E\left(X_i^2\right)\right]^2 = 3 - 1 = 2,$$

由于 $X_1, X_2, \cdots, X_n$ 相互独立, 所以 $X_1^2, X_2^2, \cdots, X_n^2$ 也相互独立, 于是

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

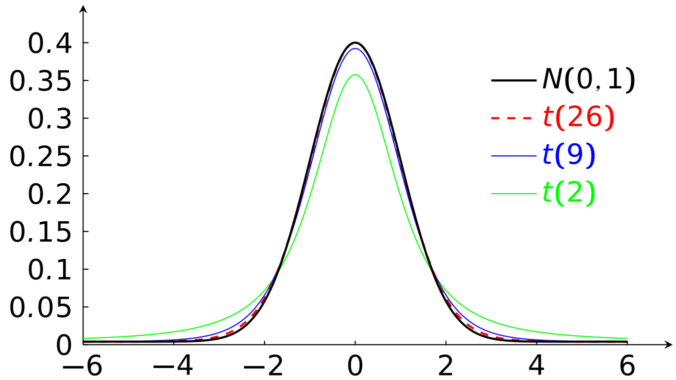
定义 **2.** 设有分布函数 $F(x)$, 对给定的 $\alpha(0 < \alpha < 1)$, 若有

$$P\{X > x_\alpha\} = \alpha,$$

则称点 x_α 为 $F(x)$ 的上 α 分位点.

当 $F(x)$ 有概率密度 $f(x)$ 时, 上式可写成

$$P\{X > x_\alpha\} = \int_{x_\alpha}^{+\infty} f(x)\mathrm{d}x = \alpha.$$



称满足 $F(x_\alpha) = \alpha$ 的 x_α 为 F 的下 α 分位点.

定义 3. 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 1. 设 $\alpha = 0.05$，$n = 20$，查表得

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

定义 4. 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 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.

注记. t 分布与标准正态分布的关系：$t(\infty) = N(0,1)$.

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 **2.** $t_{0.05}(10) = 1.812, \quad Z_{0.025} = 1.960.$

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

定义 **5.** 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 **3.** 自由度为 m 和 n 的 F 分布的概率密度为

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

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 $T^2 \sim F(1,n)$.

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

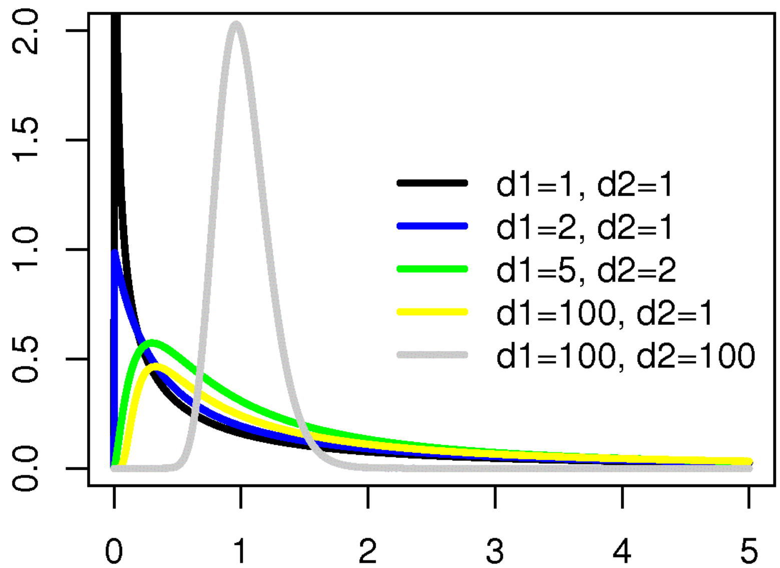


图 6.1: F 分布的密度函数

6.4.2 正态总体统计量的分布

定理 4. 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 5. 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 **6.** 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.

定理 **7.** 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

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