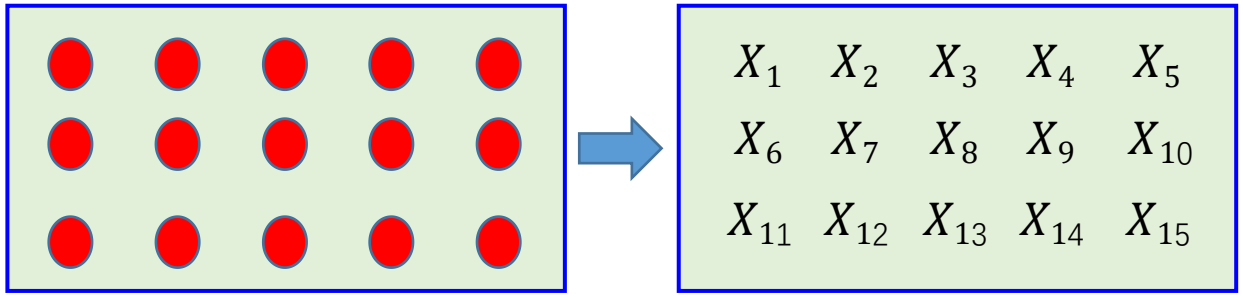
抽样分布

第一节：基本概念

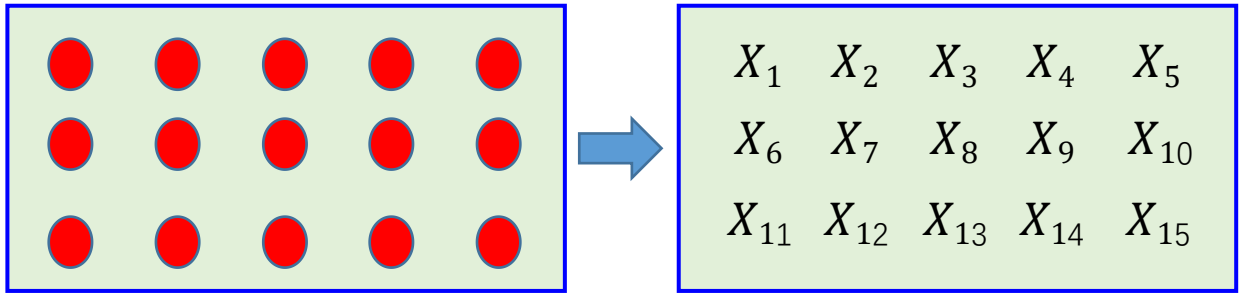
- 胡政发，肖海霞，应用数理统计与随机过程，电子工业出版社，2021年第一版
- 师义民，徐伟，秦超英，许勇，数理统计，科学出版社，2015年第四版

●总体和样本



- 总体：研究对象的全体元素组成的集合
- 个体：组成总体的每个元素
  - 例：考察某批灯泡质量时，这批灯泡的全体就组成一个总体，其中每个灯泡就是个体。
- 在实际应用中，关心的不是总体中个体的一切方面，而往往是个体的某一项或者某几项数量指标。

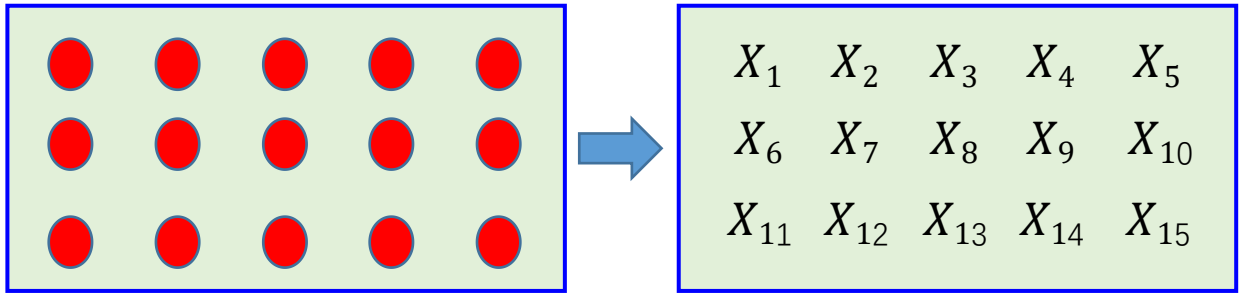
●总体和样本



➢例：考察灯泡的质量时，并不关心灯泡的形状、样式等特征，只研究灯泡的寿命、亮度等指标特征。

✓ 如果只考察灯泡寿命这一项指标时，由于一批灯泡中每个灯泡都有一个确定的寿命值，因此，自然地把这批灯泡寿命值的全体视为总体，其中每个灯泡的寿命值就是个体。

●总体和样本

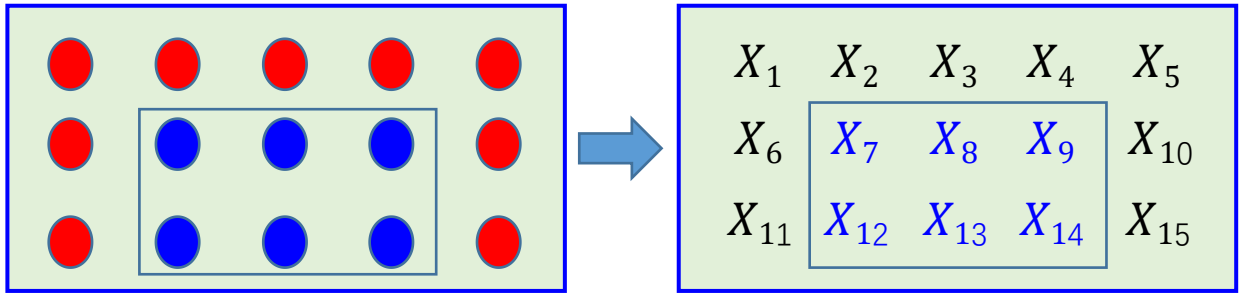


➢例：由于具有不同寿命值的灯泡的比例是按照一定规律分布的，即任取一个灯泡其寿命为某一值具有一定的概率，因而，这批灯泡的寿命是一个随机变量，也就是说，可以用一个随机变量 X 来表示这批灯泡的寿命这个总体。

□在数理统计中，任何一个总体都可以用一个随机变量来描述。

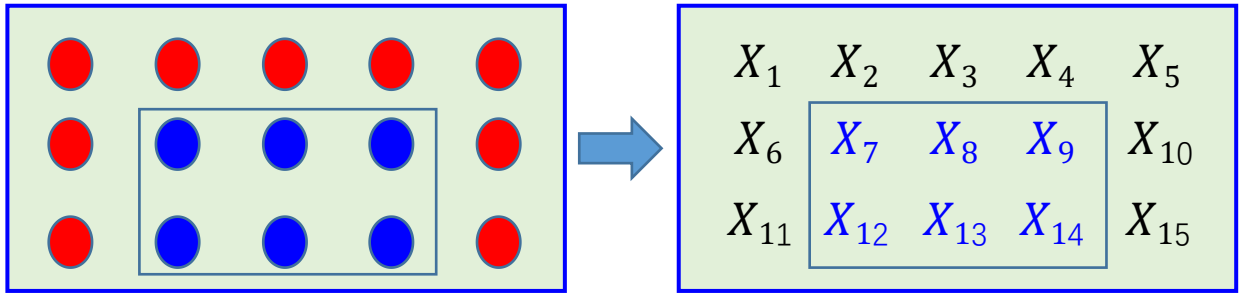
□总体的分布及数字特征，即指表示总体的随机变量的分布及数字特征，对总体的研究也就归结为对表示总体的随机变量的研究。

●总体和样本



□样本：为了解总体 X 的分布规律或者某些特征，必须对总体进行抽样观察，即从总体 X 中，随机抽取n个$X_1, X_2, \cdots, X_n$，记为$(X_1, X_2, \cdots, X_n)^T$，称它为来自总体 X 的容量为 n 的样本。

●总体和样本



- 由于每个 X_i 都是随机抽取的，它的取值就在总体 X 的可能取值范围内随机取得，自然每个X_i也是随机变量，从而样本$(X_1, X_2, \cdots, X_n)^T$是一个$n$维随机变量。在抽样观测之后，它们是$n$个数据$(x_1, x_2, \cdots, x_n)^T$，称之为样本$(X_1, X_2, \cdots, X_n)^T$的一个观测值，简称样本值。
- 样本$(X_1, X_2, \cdots, X_n)^T$可能取值的全体称为样本空间。

●总体和样本

□目的：依据从总体中抽取的一个样本值$(x_1, x_2, \cdots, x_n)^T$，对总体的分布或者某些特征进行分析判断，因而要求抽取的样本能很好地反映总体的特征且便于处理，于是，提出如下两点要求：

①代表性：要求样本$X_1, X_2, \cdots, X_n$同分布且每个X_i与总体具有相同的分布

②独立性：要求样本$X_1, X_2, \cdots, X_n$是相互独立的随机变量满足上述两条性质的样本称为简单随机样本，今后提到的样本均指简单随机样本

● 总体和样本

□ 定理1.1：设总体 X 的分布函数为$F(x)$，则来自总体 X 的样本$(X_1, X_2, \cdots, X_n)^T$的联合分布函数为$\prod_{i=1}^{n} F(x_i)$

➢例题 1.1

● 设总体X服从参数为p的两点分布，即

$$P\{X=1\}=p,\ P\{X=0\}=1-p,\ \ 0<p<1$$

试求样本$(X_1,X_2,\cdots,X_n)^T$的联合分布律

✓ 解：由于总体的分布律可以写成

$$P\{X=1\}=p,\ P\{X=0\}=1-p,\ \ 0<p<1$$

由上述定理，样本$(X_1,X_2,\cdots,X_n)^T$的联合分布律为

$$\prod_{i=1}^{n}p(x_i)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}$$

➢例题 1.2

- 设总体X服从正态分布$N(\mu,\sigma^2)$，试求样本$(X_1,X_2,\cdots,X_n)^T$的联合分布密度

✓解：

$$\prod_{i=1}^{n} f(x_i)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

●统计量和样本矩

- 样本是总体的代表和反映，但在抽取样本之后，并不能直接利用样本进行推断，而需要对样本进行加工和提炼，把样本中关于总体的信息集中起来，这便是针对不同的问题构造样本的某种函数。
- 为此，引入统计量的概念

●统计量和样本矩

□定义1.1　设$(X_1, X_2, \cdots, X_n)^T$为总体$X$的一个样本，若$f(X_1, X_2, \cdots, X_n)$为一个函数，且$f$中不含任何未知参数，则称$f(X_1, X_2, \cdots, X_n)$为一个统计量

□由于样本$(X_1, X_2, \cdots, X_n)^T$是随机变量，统计量$f(X_1, X_2, \cdots, X_n)$也是随机变量，它们应有确定的分布，统计量的概率分布称为抽样分布

●统计量和样本矩

□常用统计量——样本矩

□定义1.2 设$(X_1, X_2, \cdots, X_n)^T$是从总体$X$中抽取的样本，称统计量

样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

样本方差 $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

修正样本方差 $S_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

样本标准差 $S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

k 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

k 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

●统计量和样本矩

□用大数定律可以证明，只要总体X的k阶矩存在，则样本的k阶矩以概率收敛于总体的k阶矩，即对任意的$\varepsilon > 0$有

$$\lim_{n\to\infty} P\{|\overline{X} - \mu| < \varepsilon\} = 1$$

$$\lim_{n\to\infty} P\{|S_n^2 - \sigma^2| < \varepsilon\} = 1$$

上式中$\mu = E\{X\},\ \sigma^2 = D\{X\}$

□此结论表明，当n很大时可用一次抽样所得的样本均值和样本方差分别作为总体的均值和方差的近似值

●统计量和样本矩

□定理1.2 设总体X具有$2k$阶矩，则来自总体X的样本k阶原点矩A_k的数学期望和方差分别为

$$E\{A_k\} = \alpha_k, \quad D\{A_k\} = \frac{\alpha_{2k} - \alpha_k^2}{n}$$

其中$\alpha_k = E\{X^k\}$表示总体的k阶原点矩

■推论

$$E\{\bar{X}\} = E\{X\}, \ D\{\bar{X}\} = \frac{1}{n} D\{X\}, \ E\{S_n^2\} = \frac{n-1}{n} D\{X\}, \ E\{S_n^{*2}\} = D\{X\}$$

●经验分布函数

- 根据样本来估计和推断总体的分布函数，是数理统计要解决的一个重要问题。
- 为此，引进经验分布函数的概念，并介绍它的性质
- 设总体X的分布函数$F(X)$，现对X进行n次重复独立观测，即对总体作n次简单随机抽样，以$v_n(x)$表示随机事件$\{X \leq x\}$在这n次重复独立观测出现的次数，即n个观测值中小于等于x的个数

●经验分布函数

□对X每进行了n次重复独立观测，便得到总体的样本$(X_1, X_2, \cdots, X_n)^T$的一个观测值$(x_1, x_2, \cdots, x_n)^T$,从而对固定的$x \in (-\infty, +\infty)$可以确定$v_n(x)$所取的数值，这个数值就是$x_1, x_2, \cdots, x_n$的$n$个数中小于等于$x$的个数，若重复进行$n$次抽样，对于同一个$x$，$v_n(x)$可能取不同数值，即$v_n(x)$随样本不同样本值而取不同值，因此，$v_n(x)$是一个统计量，从而也是随机变量。

●经验分布函数

□$v_n(x)$通常称为经验频数，由于在n重独立试验中，某事件出现的次数服从二项分布，故有$v_n(x)$服从二项分布

$$P\{v_n(x)=k\}=C_n^k(P\{X\le x\})^k(1-P\{X\le x\})^{n-k}$$
$$=C_n^k(F(x))^k(1-F(x))^{n-k}$$

记为$v_n(x)\sim B(n,F(x))$

●经验分布函数

□定义1. 3 称函数$F_n(x)=\frac{v_n(x)}{n},-\infty<x<+\infty$为总体$X$的经验分布函数

□设$(X_1,X_2,\cdots,X_n)^T$是来自总体X的样本，其样本值为$(x_1,x_2,\cdots,x_n)^T$，将$x_1,x_2,\cdots,x_n$按从小到大的顺序排列并重新编号为$x_{(1)}\le x_{(2)}\le\cdots\le x_{(n)}$，则总体的经验分布函数可以表示为

$$F_n(x)=\frac{v_n(x)}{n}=\begin{cases}0, & x<x_{(1)}\\ \frac{k}{n}, & x_{(k)}\le x<x_{(k+1)},k=1,2,\cdots,n-1\\ 1, & x\ge x_{(n)}\end{cases}$$

●经验分布函数

□经验分布函数的性质

①当给定样本值$(x_1, x_2, \cdots, x_n)^T$时，$F_n(x)$是一个分布函数，满足如下性质

(i) $0 \le F_n(x) \le 1$

(ii) $F_n(-\infty) = 0,\ F_n(+\infty) = 1$

(iii) $F_n(x)$非减且右连续

●经验分布函数

□经验分布函数的性质

② $F_n(x)$是随机变量，且$nF_n(x) = v_n(x) \sim B(n, F(x))$，进而

$$E\{F_n(x)\} = F(x), \quad D\{F_n(x)\} = \frac{1}{n}F(x)[1-F(x)]$$

③当$n \to \infty$时，$F_n(x)$依概率1一致地收敛于$F(x)$, 即

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

即：当$n \to \infty$时，由样本值得到的经验分布函数$F_n(x)$是总体分布函数$F(x)$的优良估计。

抽样分布

第二节：抽样分布

- 胡政发，肖海霞，应用数理统计与随机过程，电子工业出版社，2021年第一版
- 师义民，徐伟，秦超英，许勇，数理统计，科学出版社，2015年第四版

- 抽样分布是指统计量的概率分布。确定统计量的分布是数理统计学的基本问题之一。关于统计量的分布，我们关心两类问题：

- （1）当总体X的分布已知时，对于任一自然数n，$U_n = f(X_1, X_2, \cdots, X_n)$的分布，这个分布称为统计量的精确分布。它对数理统计中的所谓小样本问题（即样本容量n 较小时的问题）的研究是很重要的；

- （2）当$n \to \infty$时，求统计量U_n的极限分布，统计量的极限分布对于数理统计中的所谓大样本问题（即样本容量n较大时的统计问题）的研究很有用处。

χ^2分布

定义1.8 设$X_1, X_2, \cdots, X_n$相互独立且同服从于标准正态分布$N(0,1)$，则称随机变量

$$\chi_n^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

服从自由度为n的χ^2分布，记为$\chi_n^2 \sim \chi^2(n)$，这里自由度n表示式（1.10）中独立变量的个数。随机变量χ_n^2称为χ^2变量。

如果平方和$\sum_{i=1}^{n} X_i^2$中，$X_1, X_2, \cdots, X_n$之间存在着k个独立的线性约束条件，则称$\sum_{i=1}^{n} X_i^2$的自由度为$n-k$（即自由变量的个数）。由于式（1.10）中，$X_1, X_2, \cdots, X_n$之间没有线性约束条件，即k=0，

χ^2分布

定理1.6. 由式（1.10）定义的随机变量χ_n^2的分布密度为

$$f(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}e^{-\frac{x}{2}}x^{\frac{n}{2}-1}, x>0\\ 0 \quad , x\le 0\end{cases} \tag{1.11}$$

其中$\Gamma(\frac{n}{2})$是伽玛函数$\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}dx$在$\alpha=\frac{n}{2}$处的值。

➢例1. 11. 设$(X_1, X_2, \cdots, X_n)^T$是来自正态总体$N(\mu, \sigma^2)$的一个样本，求随机变量

$$Y = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

的概率分布。

✓解：因为$X_1, X_2, \cdots, X_n$相互独立，且$X_i \sim N(\mu, \sigma^2)(\mathrm{i} = 1,2, \cdots, n)$. 作变换$Y_i = \frac{X_i - \mu}{\sigma}$，显然$Y_1, Y_2, \cdots, Y_n$相互独立，且$Y_i \sim N(0,1)$ $(\mathrm{i} = 1,2, \cdots, n)$.

因此由定义1. 8得：$Y = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n} Y_i^2$

服从自由度为n的χ^2分布。

- χ^2分布具有下列性质：

- 性质1 $E\chi_n^2 = \mathrm{n}, D\chi_n^2 = 2\mathrm{n}.$

- 性质2 若$\chi_1^2 \sim \chi^2(\mathrm{n}), \chi_2^2 \sim \chi^2(\mathrm{m})$,且$\chi_1^2$与$\chi_2^2$相互独立，则.

  $\chi_1^2 + \chi_2^2 \sim \chi^2(\mathrm{n}+\mathrm{m})$

- 性质2称为χ^2分布的可加性。这个性质还可以推广到多个变量的情形，即n个相互独立的χ^2变量之和亦是χ^2变量，且它的自由度等于各个χ^2变量相应自由度之和。

- 性质3 若$\chi_n^2 \sim \chi^2(\mathrm{n})$,则对任意x，有

$$\lim_{\mathrm{n}\to\infty} P\{\frac{\chi_n^2 - n}{\sqrt{2n}} \le x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

- 性质3说明χ^2变量的极限分布是正态分布，因而，当n很大时，

$\frac{\chi_n^2 - n}{\sqrt{2n}}$近似服从标准正态分布$N(0,1)$，亦即n很大时，$\chi_n^2$近似服从正态分布$N(n, 2n)$。

- 下面介绍一个比性质2更为深刻的结论——柯赫伦分解定理。

- **定理1.7（柯赫伦分解定理）**设$X_1, X_2, \cdots, X_n$相互独立，且

$X_i \sim N(0,1)\ (i = 1,2, \cdots, n)$。令$\mathrm{Q} = \sum_{i=1}^{n} X_i^2$，$Q$是自由度为$n$的$\chi^2$变量。若$Q$可以分解成

$$Q = Q_1 + Q_2 + \cdots + Q_k$$

其中$Q_i(i = 1,2, \cdots, k)$是秩为n_i的关于$(X_1, X_2, \cdots, X_n)^T$的非负二次型。则$Q_i$相互独立且 $Q_i \sim \chi^2(n_i)\ (i = 1,2, \cdots, k)$的充要条件是

$$n_1 + n_2 + \cdots + n_k = n$$

- 必要性依χ^2变量的可加性是显然的，充分性证明需要用到较多线性代数知识，故从略。

- 该定理在方差分析中起着重要的作用。它被这样应用：如果由$(X_1, X_2, \cdots, X_n)^T$构成的自由度为n的$\chi^2$变量Q能够分解成若干个关于$(X_1, X_2, \cdots, X_n)^T$的非负二次型，那么只要这若干个二次型的秩之和为n，则每个二次型均服从χ^2分布，且分布的自由度等于相应于该二次型的秩。

t分布

定义1.9 设$X\sim N(0,1), Y\sim\chi^2(n)$，且X与Y相互独立，则称随机变量

$$T=\frac{X}{\sqrt{Y/n}} \tag{1.12}$$

服从自由度为n的t分布，记为$T\sim t(n)$，随机变量T亦成为T变量。

定理1.8 由式（1.12）所定义的T变量的分布密度为

$$\varphi_T(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},-\infty<x<+\infty \tag{1.13}$$

由于式（1.13）中的函数是偶函数，所以t分布密度关于Y轴对称，而且当n很大时，t分布很接近于标准正态分布。

➢例1.12 设$X\sim N(\mu,\sigma^2),Y/\sigma^2\sim\chi^2(n)$，且X与Y相互独立。试求$T=\frac{X-\mu}{\sqrt{Y/n}}$的概率分布。

✓解 因为$X\sim N(\mu,\sigma^2)$，所以$\frac{X-\mu}{\sigma}\sim$N(0,1)，　　又$Y/\sigma^2\sim\chi^2(n)$，

由于X与Y相互独立，因此$\frac{X-\mu}{\sigma}$与Y/σ^2相互独立，从而由定义1.9，

$$\frac{X-\mu}{\sqrt{Y/n}}=\frac{(X-\mu)/\sigma}{\sqrt{(\frac{Y}{\sigma^2})/n}}\sim t(n).$$

即$T=\frac{X-\mu}{\sqrt{Y/n}}$服从自由度为n的t分布。

- t分布具有下列性质：
- 性质1 设$T\sim t(n)$，当n>2时，$ET=0$，$DT=\frac{n}{n-2}$。
- 性质2 设$T\sim t(n)$，$\varphi_T(t)$是T的分布密度，则

$$\lim_{n\to\infty}\varphi_T(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$$

- 性质2说明当$n\to\infty$时，t分布的极限分布是标准正态分布。事实上，当n>30时，t分布与标准正态分布就非常接近了。但对较小的n值，t分布与N(0,1)分布有较大的差异，而且有

$$P\{|T|\geq t_0\}\geq P\{|X|\geq t_0\}$$

$X\sim N(0,1)$，即在t分布的尾部比标准正态分布尾部有更大的概率。

F分布

定义1.10 设 $X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$，且X与Y相互独立，则称随机变量

$$F=\frac{X/n_1}{Y/n_2} \tag{1.14}$$

服从自由度为（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。

定理1.9 自由度为(n_1,n_2)的F分布的分布密度为

$$\varphi_F(x)=\begin{cases}\dfrac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\dfrac{n_1}{n_2}\right)\left(\dfrac{n_1}{n_2}x\right)^{\frac{n_1}{2}-1}\left(1+\dfrac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, & x>0\\ 0, & x\le 0\end{cases} \tag{1.15}$$

➢例1.13 已知$T\sim t(n)$，证明：$T^2\sim F(1,n)$。

✓证明 因为$T\sim t(n)$，由定义1.9有$T=\frac{X}{\sqrt{Y/n}}$，其中

$X\sim N(0,1), Y\sim\chi^2(n)$,X与Y独立，那么

$$T^2=\frac{X^2}{Y/n},$$

由于$X^2\sim\chi^2(1)$，且X^2与Y相互独立，由定义1.10有

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

F分布具有下列性质：

性质1 设$F\sim F(n_1,n_2)$, 则

$$E\{F\}=\frac{n_2}{n_2-2}(n_2>2) \qquad D\{F\}=\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}(n_2>4)$$

性质2 设$F\sim F(n_1,n_2)$，则$\frac{1}{F}=\frac{Y/n_2}{X/n_1}\sim F(n_2,n_1)$

性质3 设$F\sim F(n_1,n_2)$，则当$n_2>4$时，对任意x有

$$\lim_{n_1\to\infty}P\left\{\frac{F-E\{F\}}{\sqrt{D\{F\}}}\le x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

性质3说明，当n_1充分大且$n_2>4$时，自由度为(n_1,n_2)的F分布近似服从正态分布

- 下面的定理在方差分析中起重要作用。
- 定理1. 10 设$X_1, X_2, \cdots, X_n$相互独立，且同服从$N(0, \sigma^2)$分布，

$Q_i(\mathrm{i} = 1,2, \cdots, \mathrm{k})$是关于$(X_1, X_2, \cdots, X_n)^T$的秩（即自由度）为$n_i$的非负二次型，且

$$Q_1 + Q_2 + \cdots + Q_k = \sum_{i=1}^{n} X_i^2$$

$$n_1 + n_2 + \cdots + n_k = n$$

则

$$F_{ij} = \frac{Q_i}{Q_j}\frac{n_j}{n_i}$$

服从自由度为(n_i, n_j)的F分布。

概率分布的分位数

定义1.11 设 X是随机变量，对于给定的实数$\alpha(0<\alpha<1)$，若存在x_α使

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

则称x_α为X（或它的概率分布）的上侧分位数。

如果$X\sim N(0,1)$,将标准正态分布的上侧分位数记为u_α，它满足关系式 $P\{X>u_\alpha\}=1-P\{X\le u_\alpha\}=1-\Phi(u_\alpha)=\alpha$,

即 $\Phi(u_\alpha)=1-\alpha$。给定α，查附表1可得u_α，如$u_{0.05}=1.645$ $u_{0.025}=1.96$等。

由于标准正态分布的对称性，显然有 $u_\alpha=-u_{1-\alpha}$