数理统计

数理统计是具有广泛应用的一个数学分支，它以概率论为理论基础，根据试验或观察得到的数据，来研究随机现象，对研究对象的客观规律性作出种种合理的估计和判断.

数理统计的内容包括:

如何收集、整理数据资料;

如何对所得的数据资料进行分析和研究,从而对所研究的对象的性质、特点作出推断.

假定某市成年男性的身高服从正态分布，希望得到平均身高μ：

在数理统计中，不是对所研究的对象全体(称为总体)进行观察，而是抽取其中的部分(称为样本)进行观察获得数据（抽样），并通过这些数据对总体进行推断.

假定某市成年男性的身高服从正态分布，希望得到平均身高μ：

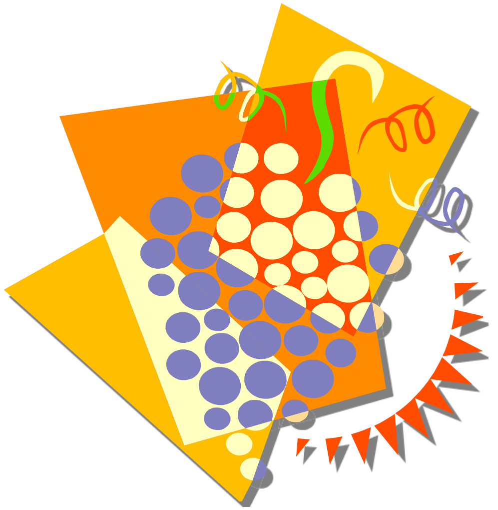
- μ的大小如何；
- μ大概落在什么范围内；
- 能否认为某一说法成立（如 $\mu \le 1.68$）.

第六章　样本及抽样分布

第一节　　随机样本

第二节　　直方图和箱线图

第三节　　抽样分布

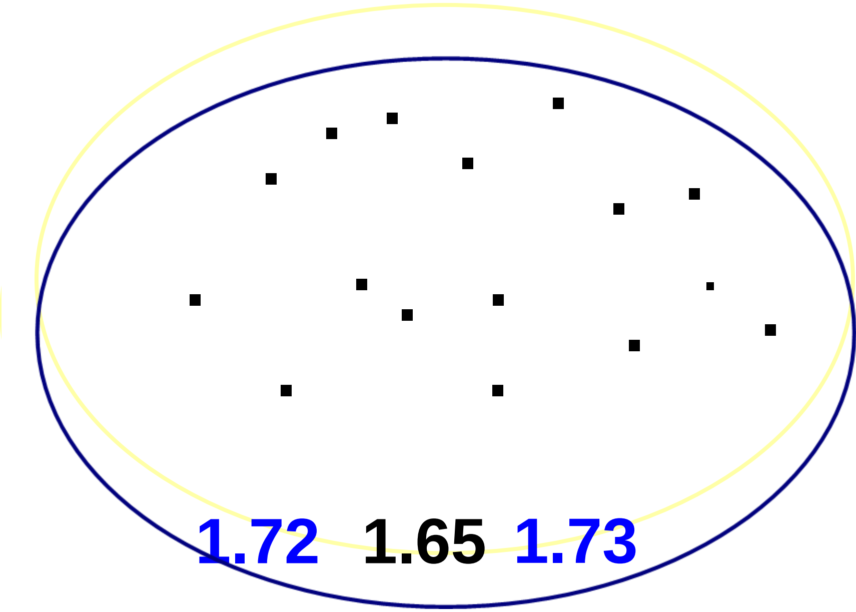


总体：研究对象的全体

个体：每个对象

总体：研究对象的全体

个体：每个对象



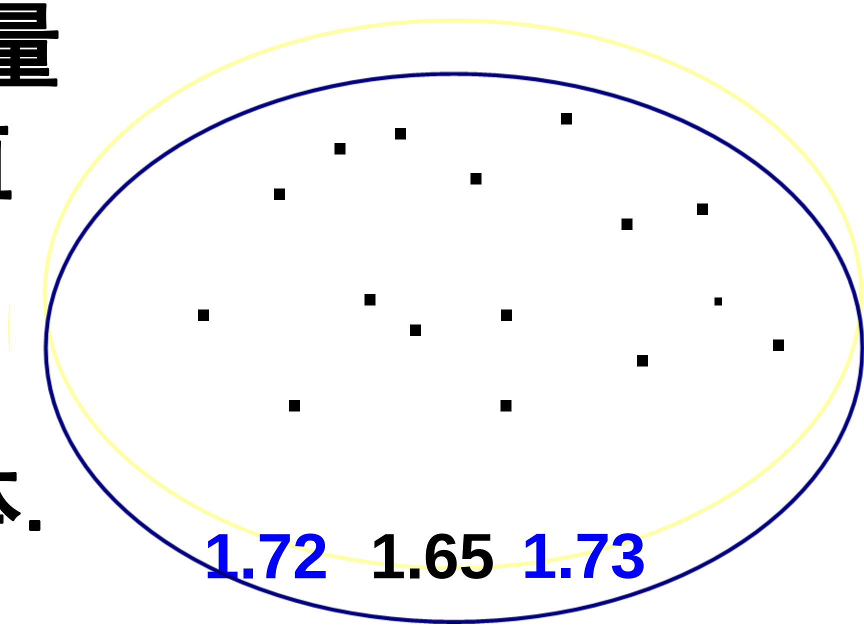
总体：研究对象的全体

个体：每个对象

总体:研究对象的某项数量指标的全部可能的观察值

个体:

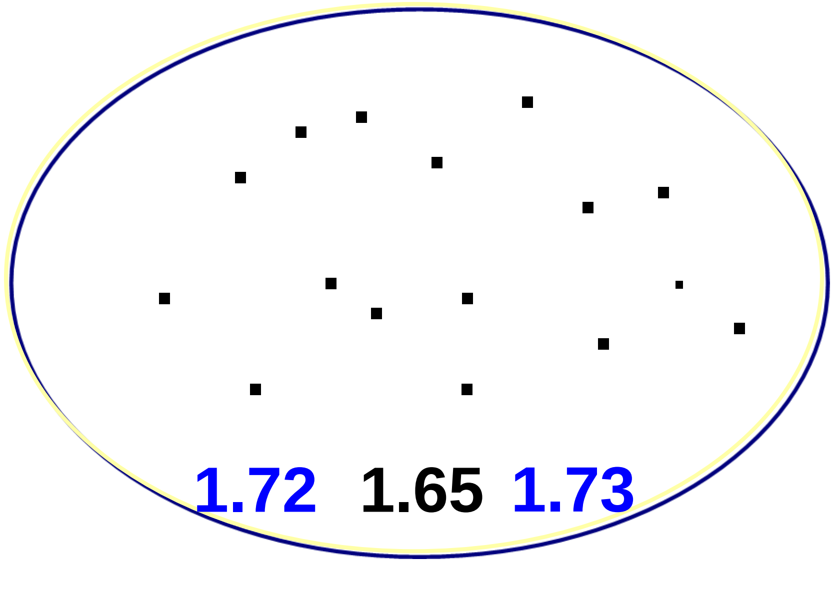
每一个可能观察值为个体.



总体中所包含的个体的个数称为总体的容量.

容量为有限的称为有限总体;
容量为无限的称为无限总体.

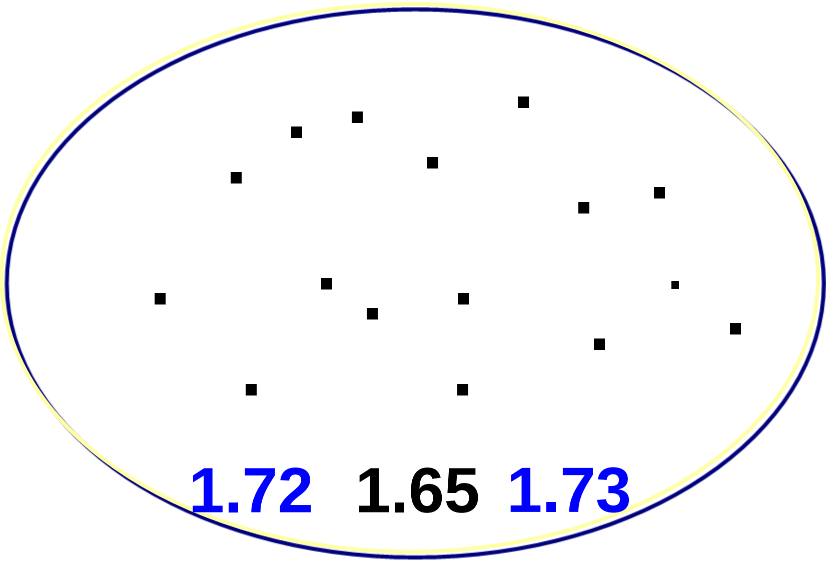
◆从总体中任取一个个体，以X表示结果.



◆从总体中任取一个个体，以X表示结果.

◆随机变量X的所有可能取值就是总体中的数值;

◆X的取值规律就是总体中数值的规律.



例 从0,1,2中随机抽取一个数，用X表示抽取结果,可以得到X的分布律

例 从0,1,2中随机抽取一个数，用X表示抽取结果,可以得到X的分布律

X	0	1	2
P	1/3	1/3	1/3

例 从0,1,2中随机抽取一个数，用*X*表示抽取结果,可以得到*X*的分布律

X	0	1	2
P	1/3	1/3	1/3

- ***X*的所有可能取值就是总体中的数值;**
- ***X*的取值规律就是总体中数值的规律.**

例　从0,1,2,2中随机抽取一个数，用X表示抽取结果,可以得到X的分布律

例　从0,1,2,2中随机抽取一个数，用X表示抽取结果,可以得到X的分布律

X	0	1	2
P	1/4	1/4	1/2

例 从0,1,2,2中随机抽取一个数，用X表示抽取结果,可以得到X的分布律

X	0	1	2
P	1/4	1/4	1/2

- **X的所有可能取值就是总体中的数值;**
- **X的取值规律就是总体中数值的规律.**

例 考察某厂的产品质量，

总体 = {该厂生产的全部合格品与不合格品}

例　考察某厂的产品质量，以0记合格品，以1记不合格品，则

总体 = {该厂生产的全部合格品与不合格品}

例　考察某厂的产品质量，以0记合格品，以1记不合格品，则

总体 = {该厂生产的全部合格品与不合格品}

= {由0或1组成的一堆数}

例　考察某厂的产品质量，以0记合格品，以1记不合格品，则

总体 = {该厂生产的全部合格品与不合格品}

= {由0或1组成的一堆数}

若以 p 表示这堆数中1的比例（不合格品率），

若从该批产品中随机抽取一件，用 X表示这一件产品的不合格数，

若从该批产品中随机抽取一件，用 X表示这一件产品的不合格数，不难看出 X 服从一个二点分布$b(1,p)$．

X	0	1
P	$1-p$	p

若从该批产品中随机抽取一件，用 X表示这一件产品的不合格数，不难看出 X 服从一个二点分布$b(1,p)$．

X	0	1
P	$1-p$	p

◆**X的所有可能取值就是总体中的数值;**

◆**X的取值规律就是总体中数值的规律.**

从总体中抽取的部分个体称为一个样本.

从总体中抽取的部分个体称为一个样本.

为调查大学生的阅读情况，某同学在图书馆抽取了部分同学进行调查。

从总体中抽取的部分个体称为一个样本.

◆从总体X中随机抽取一个个体，
以X_1表示其结果，X_1和X有相同的分布.

从总体中抽取的部分个体称为一个样本.

◆从总体X中随机抽取一个个体,

以X_1表示其结果，X_1和X有相同的分布.

◆放回，从总体X中再随机抽取一个个体,

以X_2表示其结果，X_2和X有相同的分布.

……

◆放回，从总体X中再随机抽取一个个体,

以X_n表示其结果，X_n和X有相同的分布.

从总体中抽取的部分个体称为一个样本.

◆从总体X中随机抽取一个个体，

以X_1表示其结果，X_1和X有相同的分布.

◆放回，从总体X中再随机抽取一个个体，

以X_2表示其结果，X_2和X有相同的分布.

······

◆放回，从总体X中再随机抽取一个个体，

以X_n表示其结果，X_n和X有相同的分布.

◆$X_1 \ldots X_n$为来自总体X的**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，但放回抽样使用起来不方便，当个体的总数N比要得到的样本容量n大很多时，在实际中可将不放回抽样近似地当做放回抽样来处理.

- 对于有限总体，采用放回抽样就能得到简单随机样本，但放回抽样使用起来不方便，当个体的总数N比要得到的样本容量n大很多时，在实际中可将不放回抽样近似地当做放回抽样来处理.
- 至于无限总体，因抽取一个个体不影响它的分布，所以总是用不放回抽样.

样本的两重性

- 抽取前无法预知它们的数值，因此，样本是随机变量，用大写字母 $X_1, X_2, \dots, X_n$ 表示；

样本的两重性

- 抽取前无法预知它们的数值，因此，样本是随机变量，用大写字母 $X_1, X_2, \ldots, X_n$ 表示；

- 抽取后经观测就有确定的 观测值，因此，样本又是一组数值。此时用小 写字母 $x_1, x_2, \ldots, x_n$.

P133　综合上述，给出定义

若$X_1,X_2,\ldots,X_n$为F的一个样本，
则$X_1,X_2,\ldots,X_n$相互独立，且它们的分布函数都是F，所以$(X_1,X_2,\ldots,X_n)$的分布函数为

$$F^*(x_1,x_2,\cdots x_n)=\prod_{i=1}^{n}F(x_i).$$

若$X_1,X_2,\ldots,X_n$为F的一个样本，

则$X_1,X_2,\ldots,X_n$相互独立，且它们的分布函数都是F，所以$(X_1,X_2,\ldots,X_n)$的分布函数为

$$F^*(x_1,x_2,\cdots x_n)=\prod_{i=1}^{n}F(x_i).$$

若X具有概率密度f，则$(X_1,X_2,\ldots,X_n)$的概率密度为

$$f^*(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)$$

第三节 抽样分布

由样本去推断总体情况，需要对样本进行“加工”，这就要构造一些样本的函数，它把样本 中所含的（某一方面）的信息集中起来.

定义 设$X_1,X_2,\ldots,X_n$是来自总体X的一个样本, $g(X_1,X_2,\ldots,X_n)$是$X_1,X_2,\ldots,X_n$的函数, 若g中不含未知参数, 则称$g(X_1,X_2,\ldots,X_n)$是一**统计量**.

定义 设$X_1,X_2,\ldots,X_n$是来自总体X的一个样本, $g(X_1,X_2,\ldots,X_n)$是$X_1,X_2,\ldots,X_n$的函数, 若g中不含未知参数, 则称$g(X_1,X_2,\ldots,X_n)$是一***统计量***.

$$X_1,X_2,\ldots,X_n \xrightarrow{\text{观察值}} x_1,x_2,\ldots,x_n$$

$$g(X_1,X_2,\ldots,X_n) \xrightarrow{\text{观察值}} g(x_1,x_2,\ldots,x_n)$$

思考

设$X_1,\cdots X_n$ 为来自总体 $X \sim N(\mu,\sigma^2)$ 的一个样本，其中μ未知,σ^2已知，问下列哪些是统计量?

$$\frac{X_1+X_n}{2};\qquad \frac{X_1+\cdots+X_n}{n}-\mu\ ;$$

$$\frac{(X_1+X_n)^2}{\sigma^2};$$

设$X_1,\cdots X_n$ 为来自总体X 的一个样本，

数理统计中最常用的统计量及其观察值有：

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

观察值记为 $$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad (2)$$

设$X_1,\cdots X_n$ 为来自总体X 的一个样本，

数理统计中最常用的统计量及其观察值有：

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

观察值记为 $$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad (2)$$

2. 样本方差 $$S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \qquad (3)$$

观察值记为 $$s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \qquad (4)$$

设$X_1,\cdots X_n$ 为来自总体X 的一个样本，

数理统计中最常用的**统计量**及其**观察值**有：

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

观察值记为 $$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad (2)$$

2. 样本方差 $$S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right) \qquad (3)$$

观察值记为 $$s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right) \qquad (4)$$

3. 样本标准差 $S=\sqrt{S^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}$ (5)

它的观察值记为 $s=\sqrt{s^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}$ (6)

3. 样本标准差 $S=\sqrt{S^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}$ (5)

它的观察值记为 $s=\sqrt{s^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}$ (6)

4. 样本*k*阶原点矩 $A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k,\qquad k=1,2,\cdots$ (7)

它的观察值记为 $a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k,\qquad k=1,2,\cdots$ (8)

5. 样本k阶中心矩 $$B_k=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^k,\qquad k=1,2,\cdots \quad (9)$$

它的观察值记为 $$b_k=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^k,\qquad k=1,2,\cdots \quad (10)$$

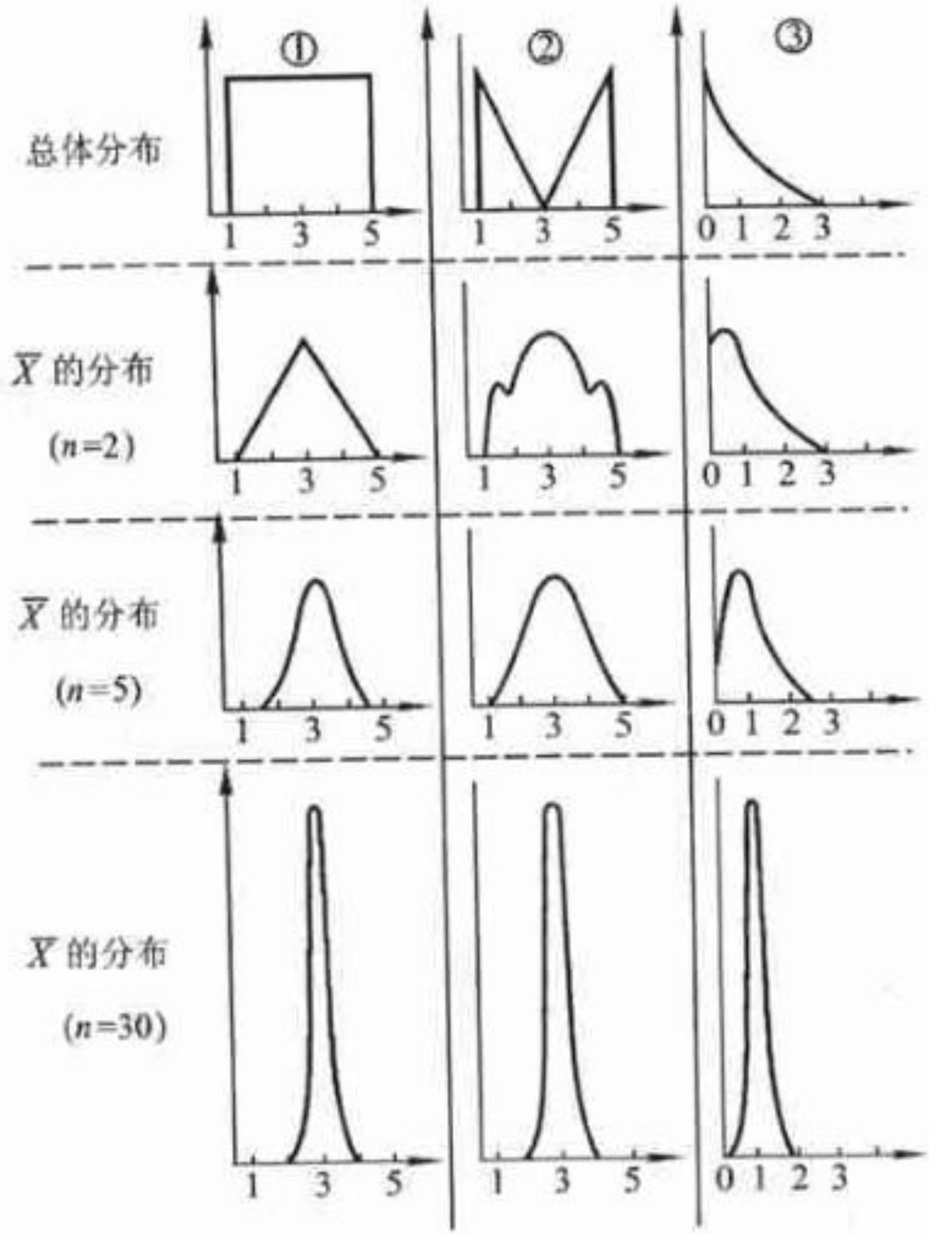
设总体X的均值为μ,方差为σ^2,

$X_1,X_2,\ldots X_n$是X的一个样本.

$$E(\overline{X})=\mu,$$

$$D(\overline{X})=\frac{\sigma^2}{n}$$

$$E(S^2)=\sigma^2$$



①
②
③
总体分布
1 3 5
1 3 5
0 1 2 3
$\overline{X}$ 的分布
(n=2)
1 3 5
1 3 5
0 1 2 3
$\overline{X}$ 的分布
(n=5)
1 3 5
1 3 5
0 1 2 3
$\overline{X}$ 的分布
(n=30)
1 3 5
1 3 5
0 1 2 3

来自正态总体的几个常用统计量的分布

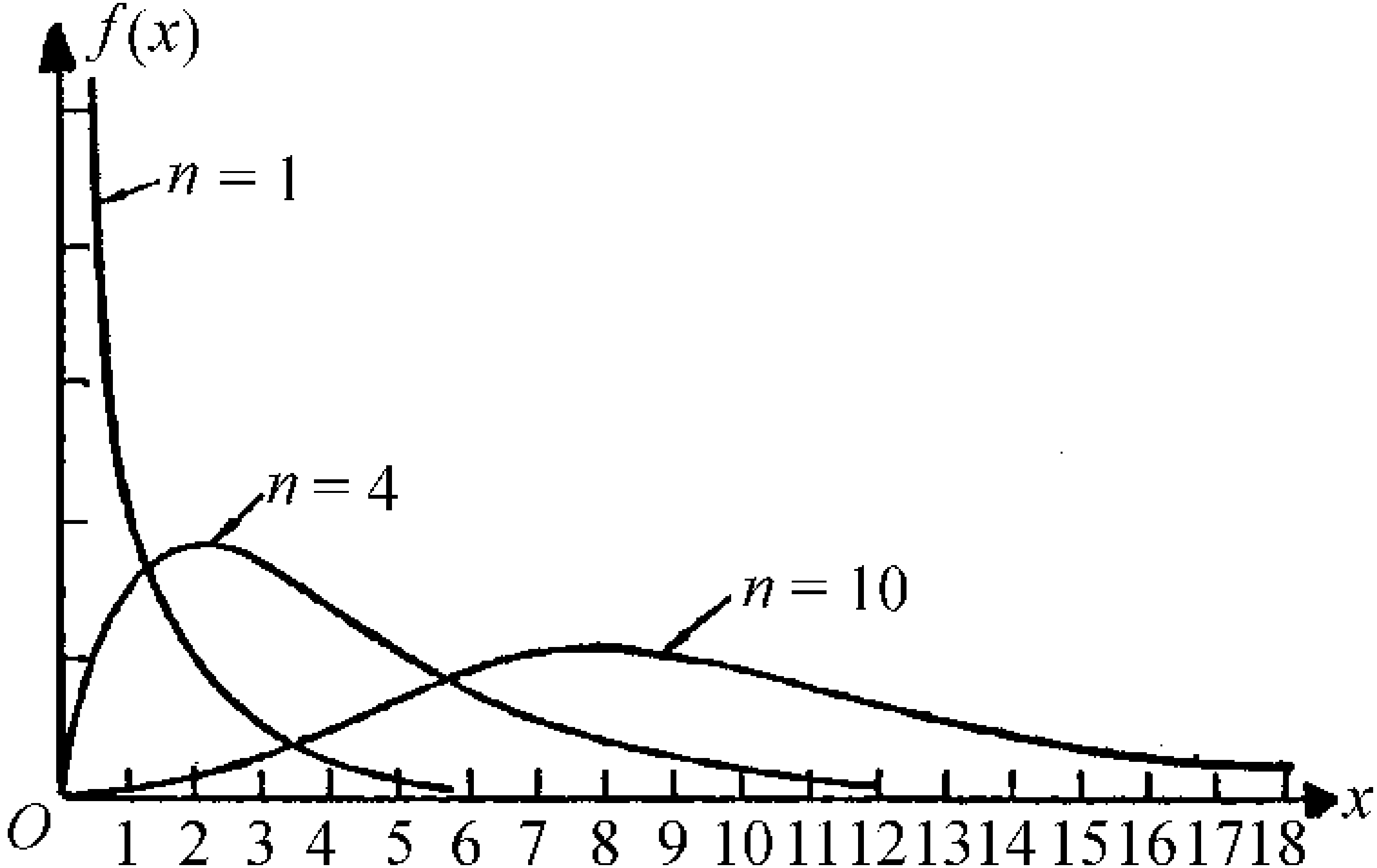
来自正态总体的几个常用统计量的分布

（一） χ^2分布

$X_1, X_2, \ldots X_n$ 是来自总体$N(0,1)$的样本, 则称统计量

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

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



f(x)
n = 1
n = 4
n = 10
O
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
x

若总体$X \sim N(0,1)$，从此总体中取一个容量为3的样本X_1, X_2，X_3,设

$$Y=(X_1+X_2+X_3)^2$$

试决定常数C，使随机变量CY服从χ^2分布.

χ^2分布的可加性

设$\chi_1^2 \sim \chi^2(n_1)$, $\chi_2^2 \sim \chi^2(n_2)$, 并且$\chi_1^2$, χ_2^2独立,

则有

$$\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2).$$

3. 若$X \sim \chi^2(n)$，则

$$E(X)=n,\ D(X)=2n.$$

事实上，由$X_i \sim N(0,1)$, 故$E(X_i^2) = D(X_i) = 1$

$$D(X_i^2) = E(X_i^4) - [E(X_i^2)]^2 = 3 - 1 = 2$$

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

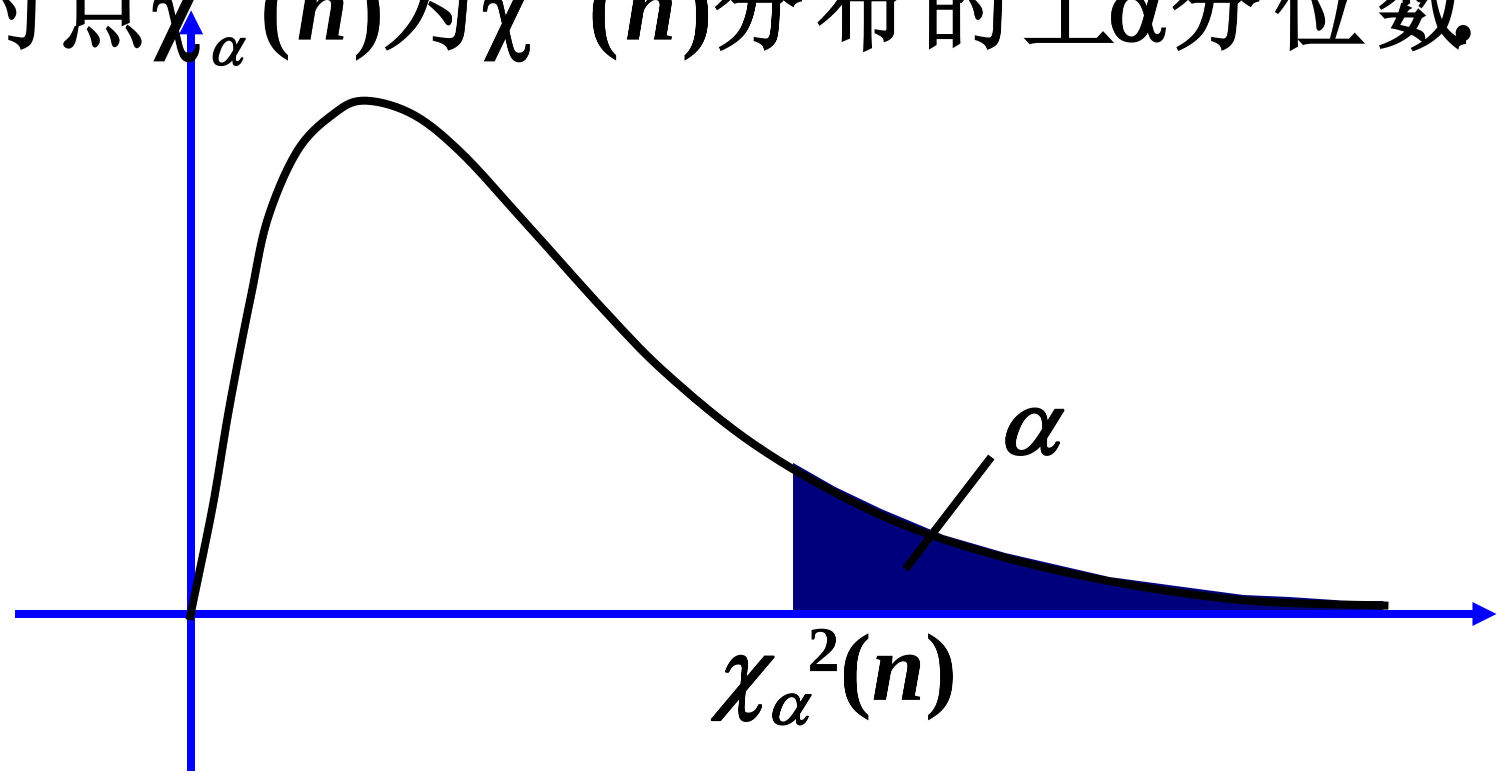
χ^2分布的分位数

χ^2分布的分位数

对于给定的正数α, 0<α<1，称满足

$$P\{\chi^2 > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2(n)}^{\infty} f(y)\,\mathrm{d}\,y = \alpha$$

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



（二） t分布（学生氏分布）

（二） t分布

设 $X \sim N(0,1)$, $Y \sim \chi^2(n)$, 且X,Y相互独立,称

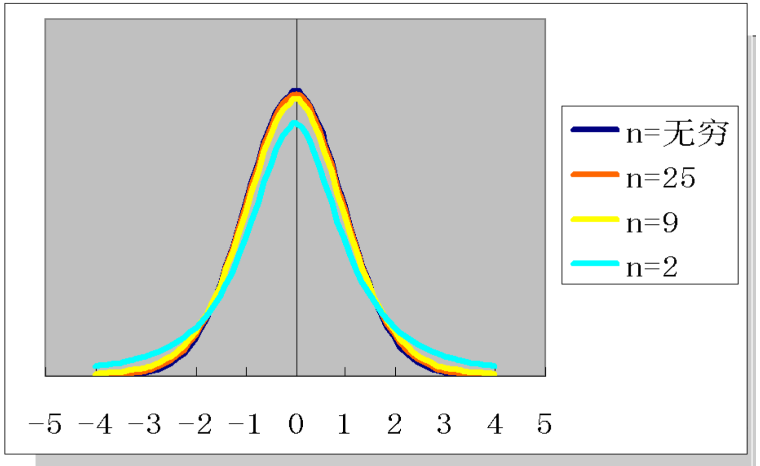
$$t = \frac{X}{\sqrt{Y/n}}$$

服从自由度为n的t分布.记为$t \sim t(n)$.

t分布的性质：

t分布的密度函数关于$t=0$对称.当n充分大时，

其图形近似于标准正态分布概率密度的图形，



t分布的概率密度曲线

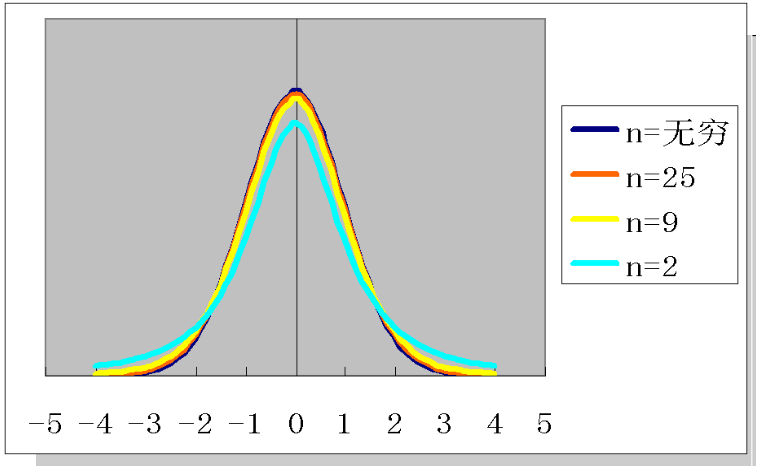
t分布的性质：

t分布的密度函数关于$t=0$对称.当n充分大时，

其图形近似于标准正态分布概率密度的图形，

再由Γ函数的性质有 $$\lim_{n\to\infty} h(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

即当n足够大时，$t \overset{\text{近似}}{\sim} N(0,1)$.

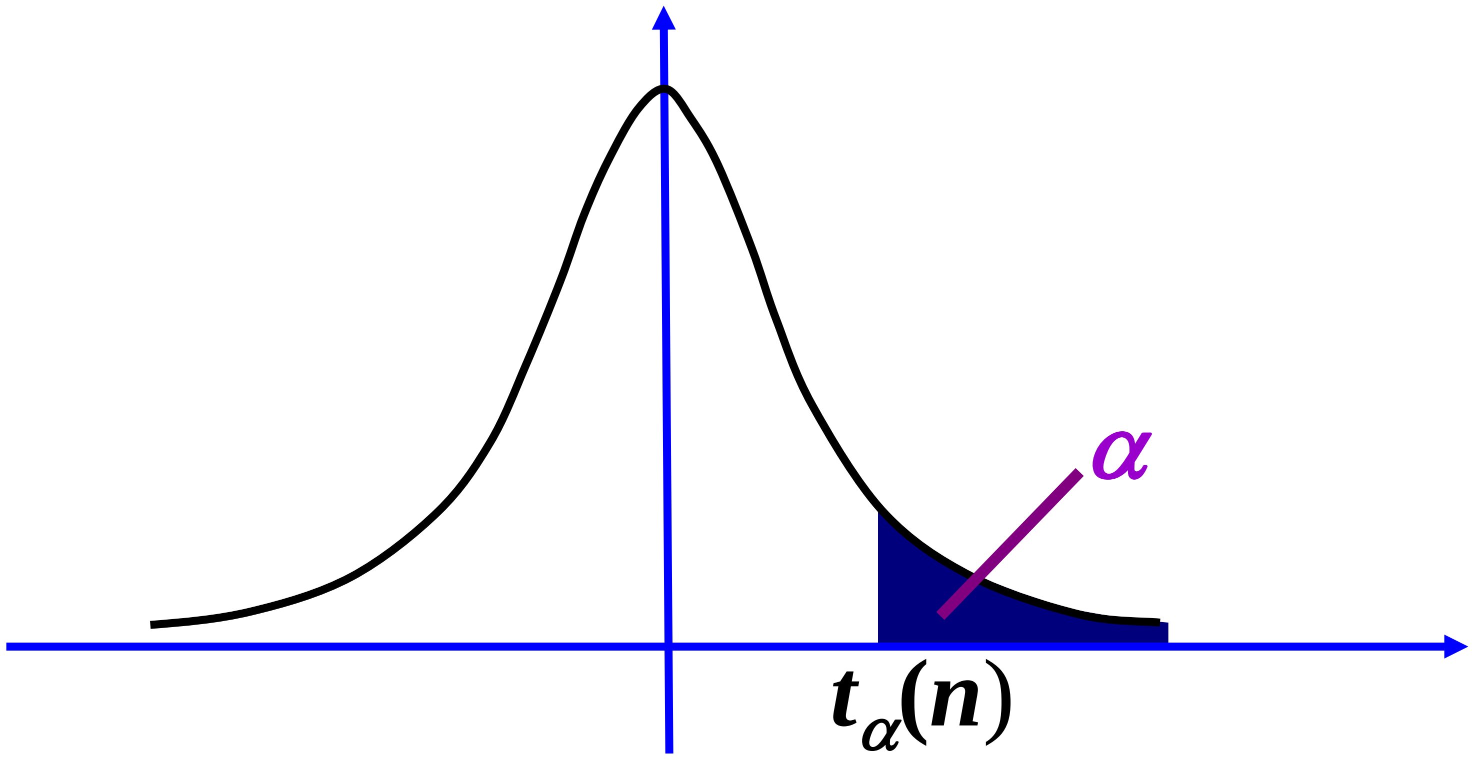


t分布的概率密度曲线

***t*分布的分位数** 对于给定的α, 0<α<1,

称满足条件 $P\{t > t_\alpha(n)\} = \int_{t_\alpha(n)}^{\infty} h(t)dt = \alpha$

的点$t_\alpha(n)$为$t(n)$分布的上α分位数.



(三) F分布

设 $U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, 且U,V相互独立,

称随机变量 $$F = \frac{U/n_1}{V/n_2}$$

服从自由度为(n_1,n_2)的F分布.记为 $F \sim F(n_1,n_2)$.

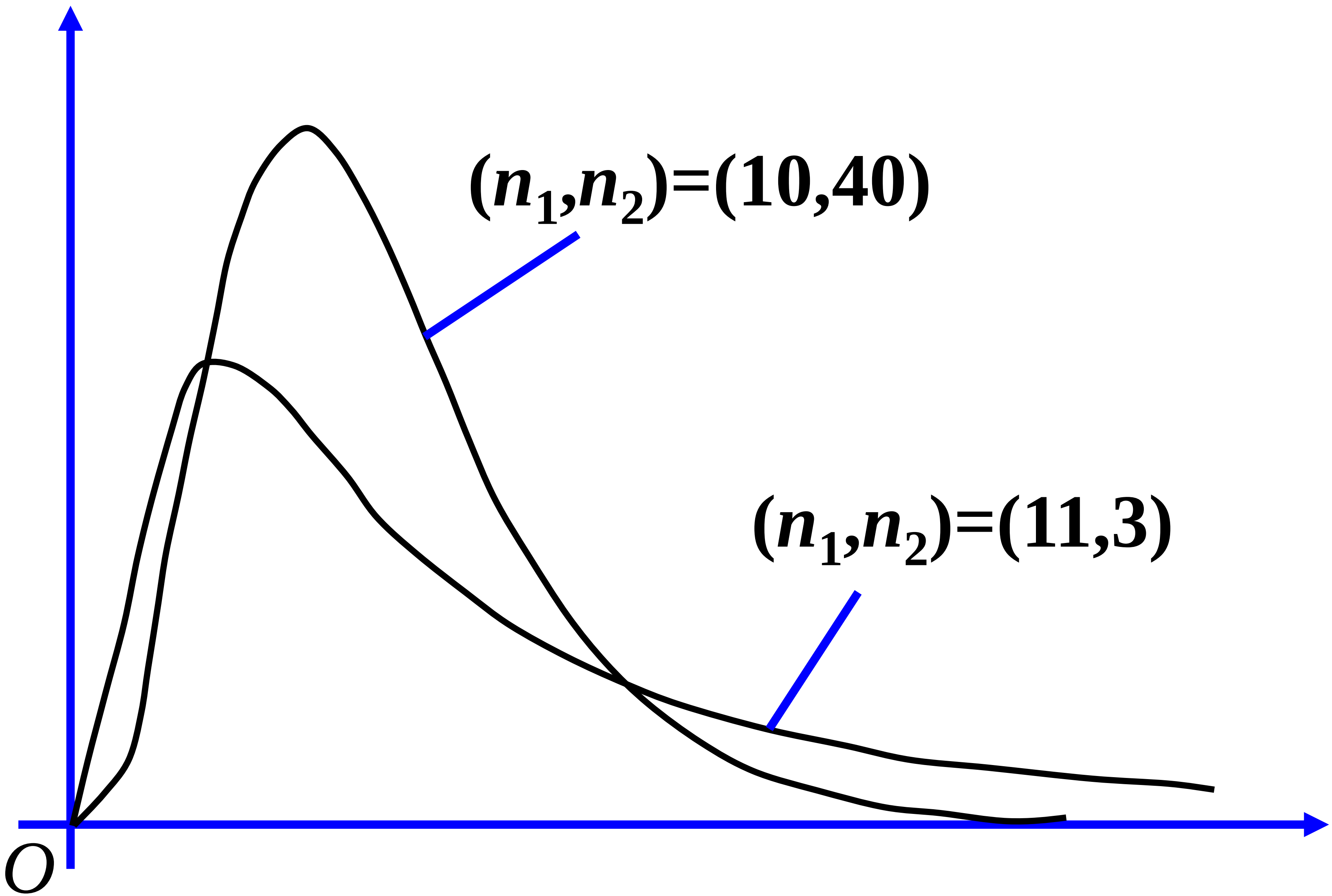
(三) F分布

设 $U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, 且U,V相互独立,

称随机变量 $$F=\frac{U/n_1}{V/n_2}$$

服从自由度为(n_1,n_2)的F分布.记为 $F \sim F(n_1,n_2)$.

若 $F \sim F(n_1,n_2)$,则 $1/F \sim F(n_2,n_1)$.



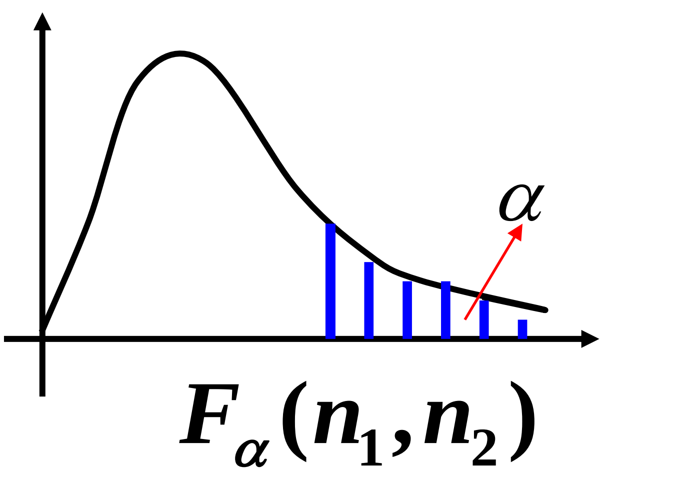
$(n_1,n_2)=(10,40)$
$(n_1,n_2)=(11,3)$
O

F分布的分位数

对于给定的$\alpha(0<\alpha<1)$，称满足条件

$$P\{F>F_{\alpha}(n_1,n_2)\}=\int_{F_{\alpha}(n_1,n_2)}^{\infty}\psi(y)\,\mathrm{d}\,y=\alpha$$

的点$F_{\alpha}(n_1,n_2)$为F分布的上α分位数。



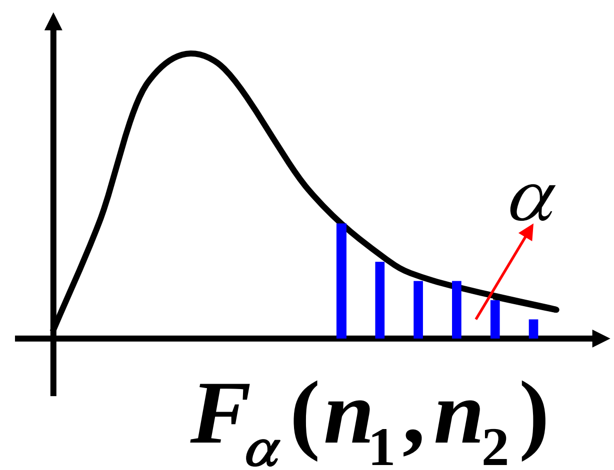
F分布的分位数

对于给定的$\alpha(0<\alpha<1)$，称满足条件

$$P\{F>F_\alpha(n_1,n_2)\}=\int_{F_\alpha(n_1,n_2)}^{\infty}\psi(y)\,\mathrm{d}\,y=\alpha$$

的点$F_\alpha(n_1,n_2)$为F分布的<u>上α分位数</u>。

$$F_{1-\alpha}(n_1,n_2)=1/F_\alpha(n_2,n_1)$$



$z_{0.025}=$__________; $z_{0.95}=$__________;

$\chi^2_{0.05}(3)=$________; $\chi^2_{0.95}(3)=$________;

$t_{0.05}(5)=$__________; $t_{0.975}(5)=$__________;

$F_{0.025}(10,10)=$__________; $F_{0.95}(8,10)=$__________;

来自正态总体的几个常用统计量的分布

（一） χ^2分布

（二） t 分布

(三) F 分布

（四）正态总体的样本均值与样本方差的分布

设产品的某质量指标 $X \sim N(\mu, \sigma^2)$

考虑总体均值μ、总体方差σ^2的统计推断问题

设总体X的均值为μ,方差为σ^2,

$X_1, X_2, \ldots X_n$是X的一个样本.

$$E(\overline{X}) = \mu,$$

$$D(\overline{X}) = \frac{\sigma^2}{n}$$

$$E(S^2) = \sigma^2$$

定理一　设$X_1,X_2,\ldots,X_n$是来自总体$N(\mu,\sigma^2)$的样本，$\overline{X}$是样本均值, 则有

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

定理二 设$X_1,X_2,\ldots,X_n$是来自总体$N(\mu,\sigma^2)$的样本，$\overline{X}$是样本均值, 则有

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

定理三 设$X_1,X_2,\ldots,X_n$是来自总体$N(\mu,\sigma^2)$的样本，$\overline{X}$和S^2是样本均值和样本方差, 则有

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

(2) $\overline{X}$与S^2独立.

定理四

设$X_1,X_2,\ldots,X_n$是取自正态总体 $N(\mu,\sigma^2)$ 的样本，$\overline{X}$和S^2分别为样本均值和样本方差，则有

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

设甲厂产品的某质量指标 $X \sim N(\mu_1, \sigma_1^2)$

设乙厂产品的某质量指标 $Y \sim N(\mu_2, \sigma_2^2)$

为了比较产品质量指标，需要考虑

$\mu_1 - \mu_2$, σ_1^2 / σ_2^2 的统计推断问题。

设产品的某质量指标$X \sim N(\mu_1, \sigma_1^2)$,

由于原材料的改变、或设备条件发生变化、或技术革新等因素的影响，使得产品质量指标可能发生变化，此时产品的质量指标为$Y \sim N(\mu_2, \sigma_2^2)$.

为了了解产品质量指标有多大的变化，需要考虑 $\mu_1 - \mu_2$，σ_1^2 / σ_2^2 的统计推断问题.

定理五（两正态总体的样本方差）

设 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$，且$X$与$Y$独立，$X_1, X_2, \ldots, X_{n_1}$是取自$X$的样本，$Y_1, Y_2, \ldots, Y_{n_2}$ 是取自Y的样本，$\overline{X}$和$\overline{Y}$分别是这两个样本的 样本均值，S_1^2和S_2^2分别是这两个样本的样本方差，则有

$$\frac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

定理五（两正态总体样本均值）

设 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, 且$X$与$Y$独立, $X_1, X_2, \ldots, X_{n_1}$是取自$X$的样本, $Y_1, Y_2, \ldots, Y_{n_2}$ 是取自Y的样本, $\overline{X}$和$\overline{Y}$分别是这两个样本的 样本均值, S_1^2和S_2^2分别是这两个样本的样本方差, 则有

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

第三节　　抽样分布

- 统计量
- 三大抽样分布
- 正态总体的样本均值与样本方差的分布

作业：

- P147　1题　4题