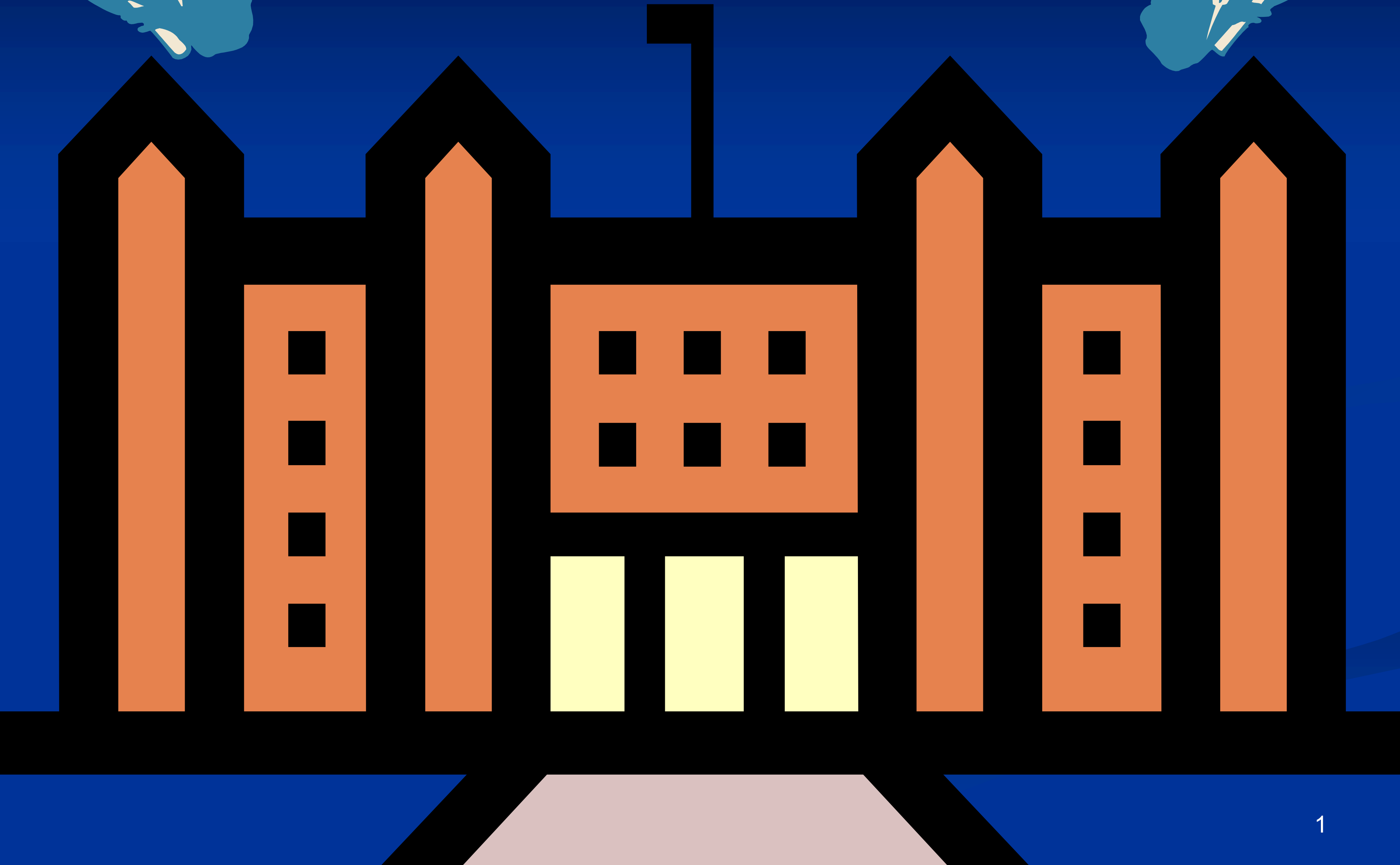
《数理统计》



在实际问题中，随机变量的分布和数字特征往往是不知道的，因此需要根据试验或观察得到的数据，来研究随机现象，对其规律作出种种合理的估计和判断。

数理统计是从局部观测资料的统计特性，来推断随机现象整体统计特性的一门科学。

要了解整体的情况，最可靠的是用普查的方法，但实际上这往往是不必要、不可能或不允许的。

学习统计无须把过多时间化在计算上，可以更有效地把时间用在基本概念、方法原理的正确理解上.国内外著名的统计软件包：*SAS*，*SPSS*，MATLAB, *STAT*等，都可以让你快速、简便地进行数据处理和分析.

数理统计学是一门应用性很强的学科．它是关于数据资料收集、整理、分析、推断,对所考察的问题作出推断和预测,直至为采取一定的决策和行动提供依据和建议的一门学科。

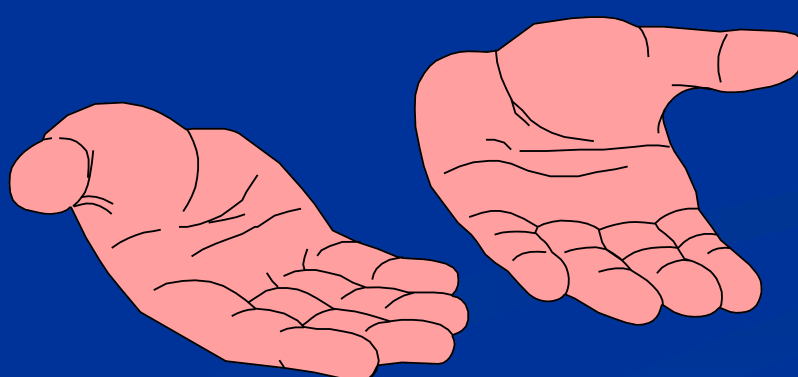
数理统计学
- 合理收集数据-试验设计、抽样调查等
- 整理分析数据-统计推断

数理统计学

- 基本概念
- 参数的估计方法
- 假设检验

3.1 样本及其抽样分布

- 总体、样本与统计量
- 常用统计量的分布



样本与统计量

一、总体与样本

1 总体 —— 研究对象全体元素组成的集合

所研究的对象的某个(或某些)数量指标的全体,它是一个随机变量(或多维随机变量).记为X.

X的分布函数和数字特征称为总体的分布函数和数字特征.

个体—— 组成总体的每个基本单元。

2. 样本：来自总体的部分个体$X_1,\cdots,X_n$,如果满足：

（1）代表性：X_i，$i=1,\ldots,n$ 与总体同分布.

（2）独立性：$X_1,\cdots,X_n$ 相互独立；

则称$X_1,\cdots,X_n$为**容量为n 的简单随机样本**,简称**样本**。而称X_1，$\cdots,X_n$的一次实现为样本观察值，记x_1，… ，x_n.

来自总体X的随机样本X_1，$\cdots$, X_n可记为

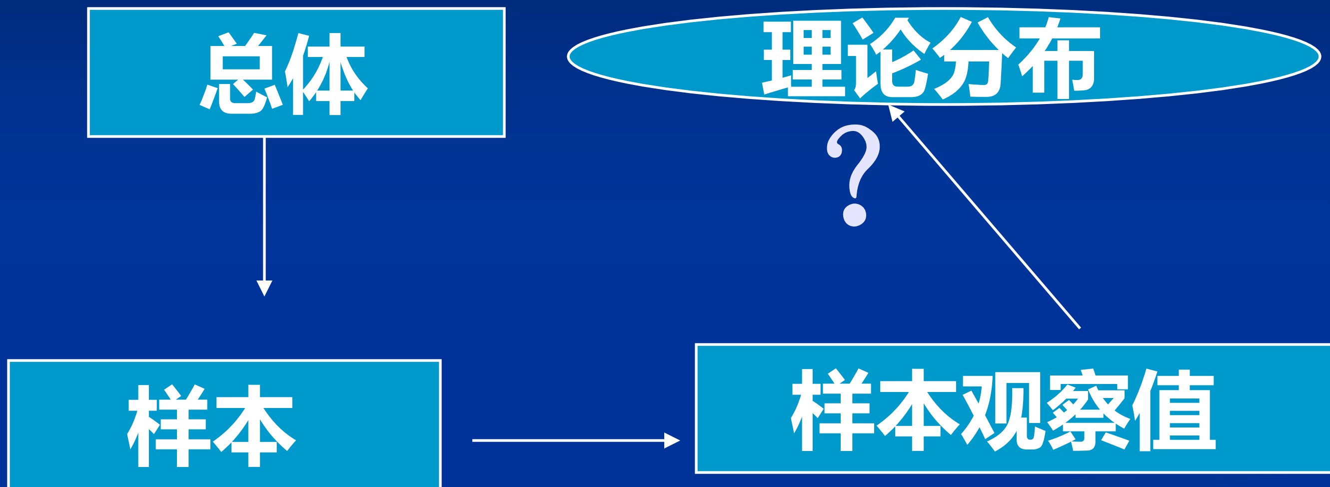
$$\boxed{X_1,\ldots,X_n \xrightarrow{i.i.d} X}$$

显然，样本联合分布函数或密度函数为

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

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

3.总体、样本、样本观察值的关系



样本空间 —— 样本所有可能取值的集合.

二、统计量

定义：样本X_1，…，X_n的函数$g(X_1，\cdots，X_n)$,如果$g(X_1，\cdots，X_n)$不含未知参数,则称$g(X_1，\cdots，X_n)$是总体X的一个统计量,记作:U

例 $X \sim N(\mu,\sigma^2)$, μ 已知 ,σ^2 为未知参数,

$(X_1,X_2,\cdots,X_n)$ 是一样本,

（1）写出样本空间与样本的密度函数；

（2）指出下列哪些是统计量？

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2,\ S'^2=\sum_{i=1}^{n}\left(X_i-\mu\right)^2/\sigma^2$$

（3）若样本观察值为1,2,3，则$\overline{X}$ 与S^2是多少?

解 $\Omega=\{(x_1,\cdots,x_n):x_i\in R, i=1,\cdots,n\}$

$$f(x_1,\cdots,x_n)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$\overline{x}=2,$$

$$s^2=\frac{1}{2}\times\{(1-2)^2+(2-2)^2+(3-2)^2\}=1$$

三、几个常用的统计量

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

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

样本均方差（标准差） $S=\sqrt{S^2},$

3.样本*k*阶矩

*k*阶原点矩 $m_k=\dfrac{1}{n}\sum_{i=1}^{n}X_i^k$

***k*阶中心矩** $M_k=\dfrac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$

性质 如果总体X的期望为μ，方差为σ^2，则

$$(1)E(\bar{X}) = E(X) = \mu \qquad (2)D(\bar{X}) = \frac{D(X)}{n} = \frac{\sigma^2}{n}$$

$$(3)\ E(S^2) = D(X) = \sigma^2$$

证明 (1)、(2)的证明留给读者，下面证明性质(3)。

$$E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$= \frac{1}{n-1} E\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$=\frac{1}{n-1}E\sum_{i=1}^{n}(X_i^2-2\bar{X}X_i+\bar{X}^2)\quad=\frac{1}{n-1}E\left[\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}\bar{X}^2\right]$$

$$=\frac{1}{n-1}E\left[\sum_{i=1}^{n}X_i^2-2\bar{X}\cdot n\bar{X}+n\bar{X}^2\right]$$

$$=\frac{1}{n-1}\left[\sum_{i=1}^{n}EX_i^2-nE(\bar{X})^2\right]$$

$$=\frac{1}{n-1}\left[\sum_{i=1}^{n}(\mu^2+\sigma^2)-n(\mu^2+\frac{1}{n}\sigma^2)\right]=\sigma^2$$

总　　结

一、总体与样本

二、统计量

三、几个常用的统计量

常用统计量的分布

确定统计量的分布是数理统计的基本问题之一

正态总体是最常见的总体, 本节介绍的几个抽样分布均对正态总体而言.

常用统计量的分布

统计学上的三大分布：

χ^2分布、t分布和F分布。

一、正态分布

定理1.若 $X_1, X_2, \cdots, X_n$ 相互独立,

$X_i \sim N(\mu_i, \sigma_i^2)$,则

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

特别地,若 $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$

则 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

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

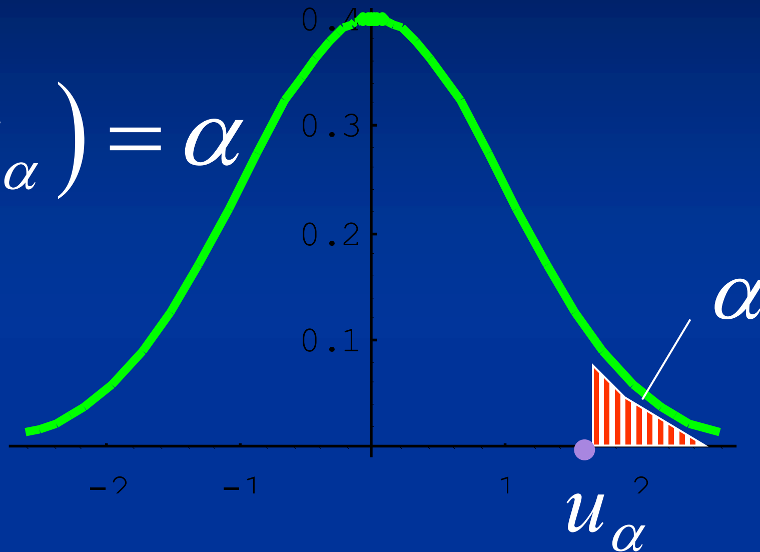
标准正态分布的 α 分位数

定义 **若** $P(X > u_\alpha) = \alpha$ 则称u_α为标准正态分布的上α 分位数.

若 $P\left(|X| > u_{\frac{\alpha}{2}}\right) = \alpha$ 则称 $u_{\alpha/2}$ 为标准正态分布的双侧α分位数.

标准正态分布的α分位数图形

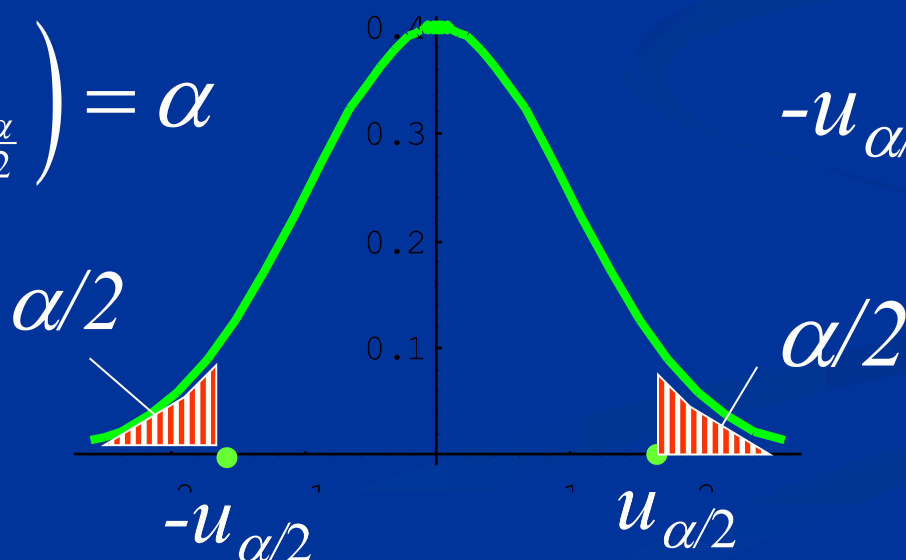
$$P\left(X > u_{\alpha}\right) = \alpha$$



$u_{0.05} = 1.645$	常用数字
$u_{0.025} = 1.96$	
$u_{0.005} = 2.575$	

$$P\left(\left| X \right| > u_{\frac{\alpha}{2}}\right) = \alpha$$

$$-u_{\alpha/2} = u_{1-\alpha/2}$$



当关注的对象的概率分布不可知，意味着只知道数据，不知道其内在规律；另一方面，关注的对象是可以分解成多种因素的组合时，就引入了抽样分布。抽样分布是描述从多个随机变量中抽取数据并且加以组合后，形成的规律。基本的抽样分布有三个：x^2（卡方）分布、F分布、t分布。

二、$\chi^2(n)$ 分布(n为自由度)

定义 设 $X_1, X_2, \cdots, X_n$相互独立,

且都服从标准正态分布N (0,1),则

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

n为自由度,为 $\sum_{i=1}^{n} X_i^2$ 中独立变量的个数。

卡方分布常用于假设检验和置信区间检验。

定理2 $X_1,\cdots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$,

(1) $\overline{X}$与S^2 相互独立；

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

例1：已知$U\sim\chi^2(10)$ 求满足

$$P\{U>\lambda_1\}=0.10, P\{U<\lambda_2\}=0.75$$

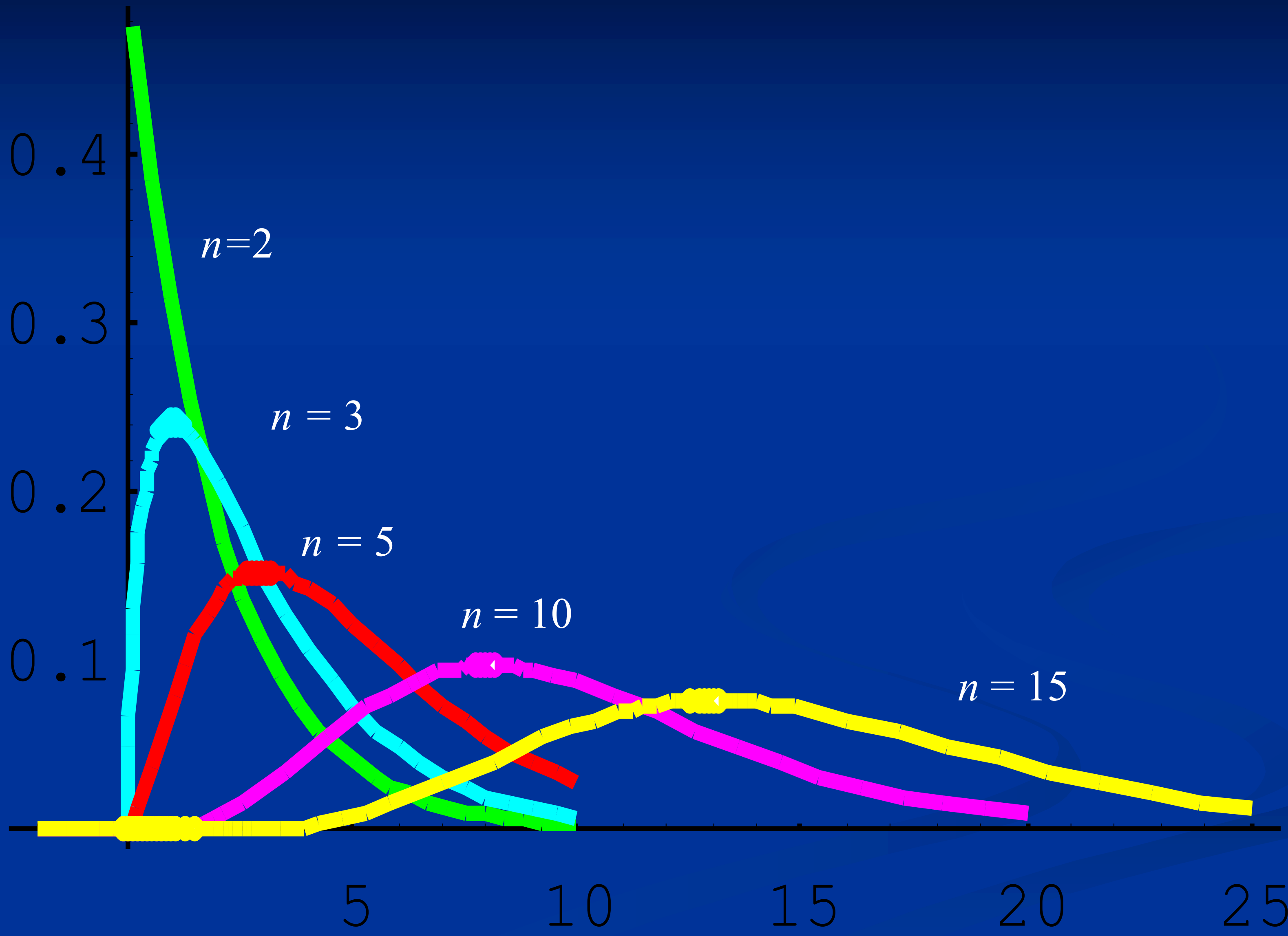
的 λ_1 和 λ_2。 查表得：$\lambda_2=12.549$

解：λ_1直接查表 $\chi^2_{0.1}(10)$ 。

而 $P\{U<\lambda_2\}=1-P\{U\geq\lambda_2\}=0.75 \quad \therefore \quad P\{U>\lambda_2\}=0.25$

x^2 分布临界值表（卡方分布）

n′	P												
	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
1	…	…	…	…	0.02	0.1	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.02	0.1	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.3	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.2	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.4	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.7	3.33	4.17	5.9	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.6	3.05	3.82	4.57	5.58	7.58	10.34	13.7	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.4	5.23	6.3	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.3
13	3.57	4.11	5.01	5.89	7.04	9.3	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.6	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25	27.49	30.58	32.8
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.3	28.85	32	34.27
17	5.7	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.6	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.2	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40



0.4
0.3
0.2
0.1
n=2
n = 3
n = 5
n = 10
n = 15
5
10
15
20
25

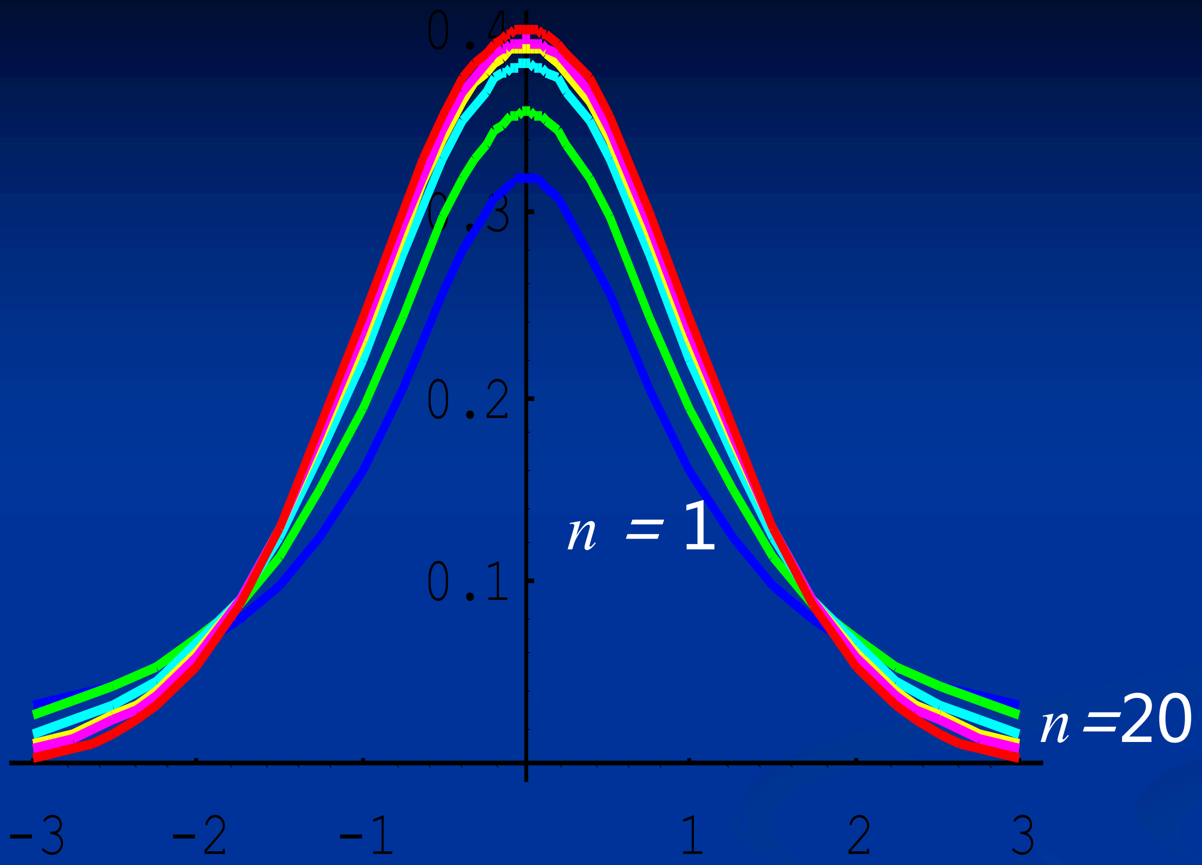
三、 t 分布 (Student 分布)

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

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

则称T服从自由度为 n 的T分布,记为$T \sim t(n)$.

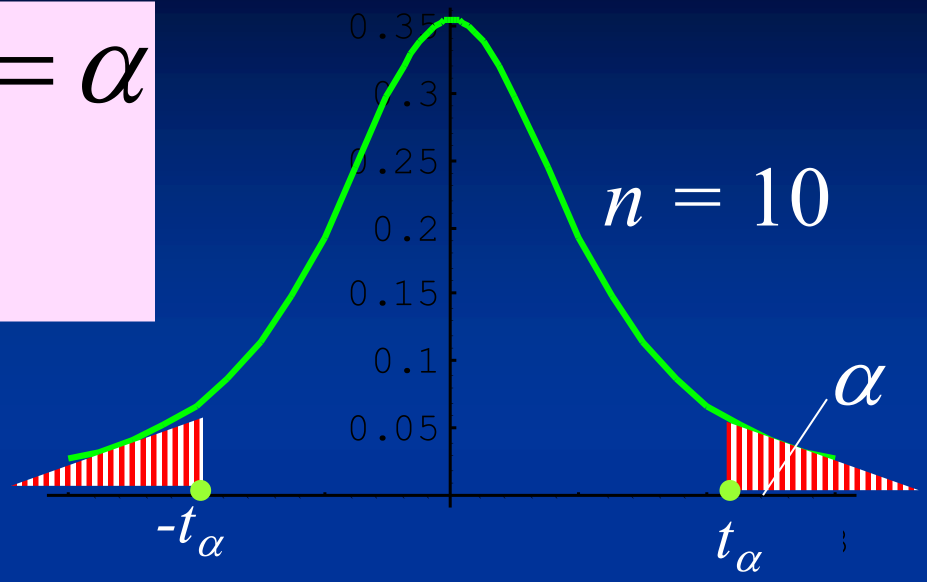
简单说一下背景，“t”，是Fisher为之取的名字。Fisher最早将这一分布命名为“Student’s distribution”，并以“t”为之标记。Student，则是William Sealy Gosset（戈塞特）的笔名。他当年在爱尔兰都柏林的一家酒厂工作，设计了一种后来被称为t检验的方法来评价酒的质量。因为行业机密，酒厂不允许他的工作内容外泄，所以当他后来将其发表到至今仍十分著名的一本杂志《Biometrika》时，就署了student的笔名。所以现在很多人知道student，知道t，却不知道Gosset。



t 分布的图形(红色的是标准正态分布)

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

$$-t_{\alpha} = t_{1-\alpha}$$



$$P\left(T > 1.8125\right) = 0.05 \Rightarrow t_{0.05}(10) = 1.8125$$

$$P\left(T < -1.8125\right) = 0.05, \quad P\left(T > -1.8125\right) = 0.95$$

$$\Rightarrow t_{0.95}(10) = -1.8125$$

定理3 $X_1,\cdots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$，则

$$T=\frac{\sqrt{n}(\bar{X}-\mu)}{S}\sim t(n-1)$$

证明

$$\overline{X}\sim N(\mu,\frac{\sigma^2}{n}) \Rightarrow \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\sim N(0,1)$$

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

$\frac{(n-1)S^2}{\sigma^2}$ 与 $\overline{X}$ 相互独立

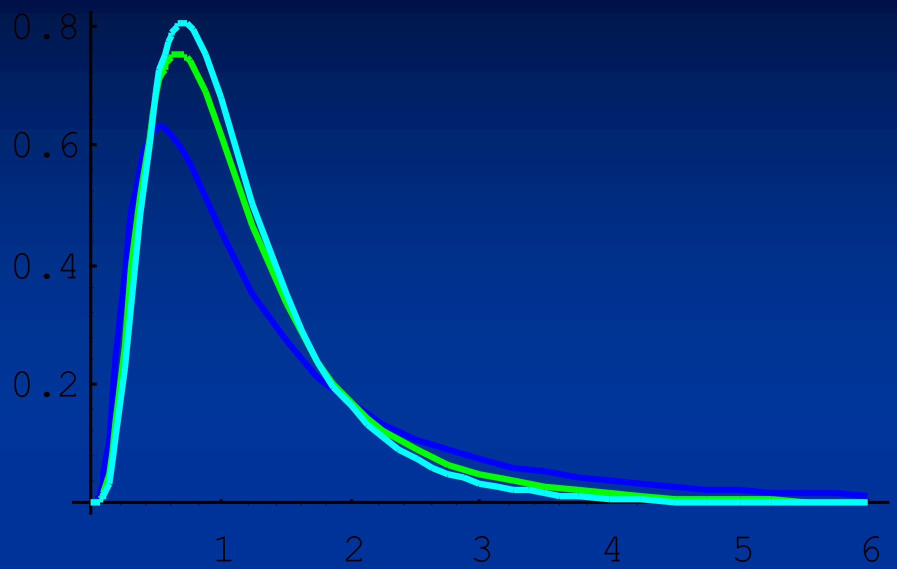
$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\Big/\frac{S}{\sigma}=\frac{\sqrt{n}(\bar{X}-\mu)}{S}\sim t(n-1)$$

四、 F 分布

定义 设 $\boxed{X \sim \chi^2(n),\ Y \sim \chi^2(m),}$ X, Y 相互独立，

令 $$\boxed{F = \frac{X/n}{Y/m}}$$

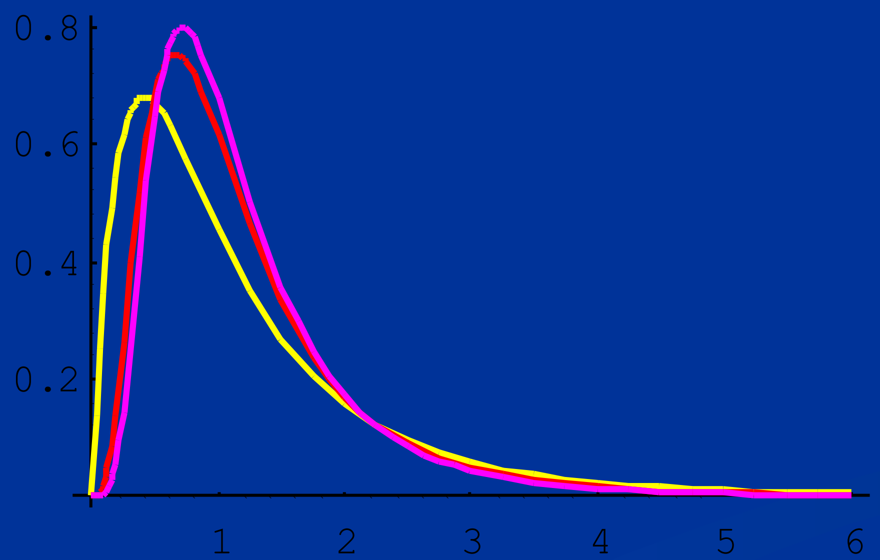
则称 F 服从为**第一自由度为n ,第二自由度为 m 的F 分布.**



$m = 10, n = 4$

$m = 10, n = 10$

$m = 10, n = 15$



$m = 4, n = 10$

$m = 10, n = 10$

$m = 15, n = 10$

总结

一、正态分布

二、$\chi^2(n)$ 分布(n为自由度)

三、 t 分布 (Student 分布)

四、 F 分布