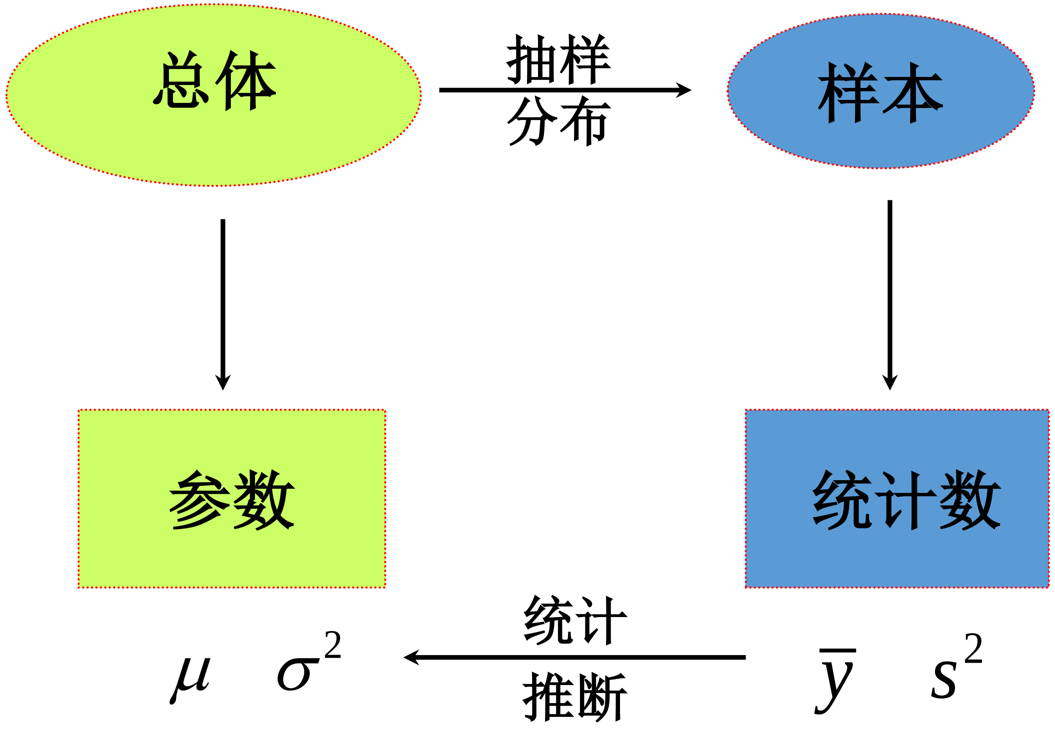
第四章 抽样分布

➢总体和样本的关系



第四章 抽样分布

4.1 随机抽样与无偏估计

（1）随机抽样

随机抽样：保证总体中的每一个个体，在每一次抽样中都有**同等的概率**被抽为样本。

总体中个体数为N，样本容量为n，所有可能样本共有N^n 个。

抽样策略：

① 假设一个较小的有限总体，进行类似无限总体的抽样模拟。

② 蒙特卡洛抽样：仅抽取一部分样本。

（2）无偏估计

例题：设有一总体N=4，具变量2，3，3，4。以样本容量n=2作独立的随机抽样，结果列于表4.1 。

$$\mu = \frac{1}{4}(2+3+3+4) = 3$$

$$\sigma^2 = \frac{1}{4}[(2-3)^2 + (3-3)^2 + (3-3)^2 + (4-3)^2] = \frac{1}{2}$$

$$\sigma = \sqrt{\frac{1}{2}} = 0.707$$

无偏估计

$$\mu = 3 \quad \sigma^2 = \frac{1}{2} \quad \sigma = 0.707$$

样本平均数$\bar{y}$的平均数：

$$\mu_{\bar{y}} = 3 = \mu$$

样本方差s^2的平均数：

$$\mu_{s^2} = \frac{1}{2} = \sigma^2$$

样本标准差s的平均数：

$$\mu_s = 0.530 \neq \sigma$$

表4.1 N=4、n=2时所有可能样本的平均数、方差和标准差

样本值	样本平均数 $\bar{y}$	样本方差s^2	样本标准差 s
2,2	2.0	0.0	0.000
2,3	2.5	0.5	0.707
2,3	2.5	0.5	0.707
2,4	3	2.0	1.414
3,2	2.5	0.5	0.707
3,3	3	0.0	0.000
3,3	3	0.0	0.000
3,4	3.5	0.5	0.707
3,2	2.5	0.5	0.707
3,3	3	0.0	0.000
3,3	3	0.0	0.000
3,4	3.5	0.5	0.707
4,2	3	2.0	1.414
4,3	3.5	0.5	0.707
4,3	3.5	0.5	0.707
4,4	4	0.0	0.000
平均数	3.0	1/2	0.530

无偏估计

扬州大学 YANGZHOU UNIVERSITY

◆ 定义

在统计上，如果所有可能样本的某一统计数的平均数等于总体的相应参数，则称该统计数为总体相应参数的无偏估值。

◆ 规律

（1）样本平均数$\bar{y}$是总体平均数μ的无偏估值。

（2）样本方差s^2是总体方差σ^2的无偏估值。

（3）样本标准差s不是总体标准差σ的无偏估值。

第四章 抽样分布

4.2 样本平均数的分布

表4.2 样本平均数的分布

$n=2$		$n=4$	
$\bar{y}$	f	$\bar{y}$	f
2.0	1	2.00	1
		2.25	8
2.5	4	2.50	28
		2.75	56
3.0	6	3.00	70
		3.25	56
3.5	4	3.50	28
		3.75	8
4.0	1	4.00	1
平均数 $\mu_{\bar{y}}$	3		3
方　差 $\sigma_{\bar{y}}^2$	1/4		1/8

N=4、n=2时所有可能样本的平均数

样本值	样本平均数
2,2	2.0
2,3	2.5
2,3	2.5
2,4	3
3,2	2.5
3,3	3
3,3	3
3,4	3.5
3,2	2.5
3,3	3
3,3	3
3,4	3.5
4,2	3
4,3	3.5
4,3	3.5
4,4	4

4.2 样本平均数的分布

◆ 性质

（1）样本平均数分布的平均数等于总体平均数

$$\mu_{\bar{y}} = \mu$$

（2）样本平均数分布的方差等于总体方差除以样本容量

$$\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}$$

进而有：

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

注： $\boldsymbol{\sigma_{\bar{y}}}$为平均数的**标准误**，是$\bar{y}$的抽样误差的度量，随着样本容量的增加而减小。

4.2 样本平均数的分布

◆ 性质

（3）如果总体服从正态分布$N(\mu,\sigma^2)$，不论样本容量n多大，样本平均数$\bar{y}$服从正态分布$N(\mu,\sigma^2/n)$。

（4）如果总体不服从正态分布，其平均数为μ，方差为σ^2，则随着样本容量n不断增大，样本平均数$\bar{y}$的分布愈来愈趋近于正态分布$N(\mu,\sigma^2/n)$。这是**中心极限定理**。

一般只要$n>30$，就可以应用这一定理，$\bar{y}$服从正态分布$N(\mu,\sigma^2/n)$。

4.2 样本平均数的分布

例题： 已知一正态总体$\mu=3$，$\sigma=0.707$，试求若样本容量$n=4$，$P\{\bar{y}<2.625\}=?$

$\bar{y} \sim N(\mu_{\bar{y}}, \sigma_{\bar{y}}^2)$ 或

$\bar{y} \sim N(\mu, \sigma^2/n)$

标准正态离差：$u=\dfrac{\bar{y}-\mu_{\bar{y}}}{\sigma_{\bar{y}}}=\dfrac{\bar{y}-\mu}{\sigma/\sqrt{n}}$

$$=\frac{2.625-3}{0.707/\sqrt{4}}$$

$$=-1.06$$

查附表1：$F(-1.06)=0.1446$

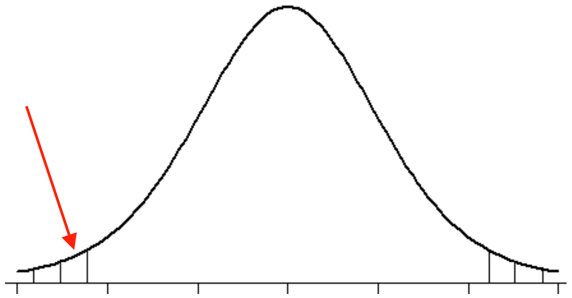
$$P\{\bar{y}<2.625\}=0.1446$$

4.2 样本平均数的分布

例题： 已知一正态总体$\mu = 29.83$, $\sigma = 1.045$,

试求若样本容量$n = 4$, $P\{|\bar{y} - \mu| > 0.5\} = ?$ $\bar{y} \sim N(\mu, \sigma^2 / n)$

$$P\{|\bar{y} - \mu| > 0.5\} = P\{\bar{y} - \mu > 0.5\} + \underline{P\{\bar{y} - \mu < -0.5\}}$$



$$u = \frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}} = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}}$$

$$= \frac{-0.5}{0.5225} = -0.96$$

查附表1：$F(-0.96) = 0.1685$ $P\{\bar{y} - \mu < -0.5\} = 0.1685$

由于正态分布左右对称，$P\{|\bar{y} - \mu| > 0.5\} = 2 \times 0.1685 = 0.3370$

第四章 抽样分布

- 4.1 随机抽样与无偏估计
- 4.2 样本平均数的分布
- **4.3 样本平均数差数的分布**
- 4.4 t分布
- 4.5 χ^2 分布
- 4.6 F分布

4.3 样本平均数差数的分布

◆ 推导

总体1 : $N_1 = 4\ (2, 3, 3, 4)$

$\mu_1 = 3 \quad \sigma_1{}^2 = 1/2 \quad n_1 = 2$

$\bar{y}_1$	f
2.0	1
2.5	4
3.0	6
3.5	4
4.0	1
总和	16

总体2 : $N_2 = 3\ (1, 2, 3)$

$\mu_2 = 2 \quad \sigma_2{}^2 = 2/3 \quad n_2 = 2$

$\bar{y}_2$	f
1.0	1
1.5	2
2.0	3
2.5	2
3.0	1
总和	9

样本平均数差数： $\bar{y}_1 - \bar{y}_2$

$16 \times 9 = 144$

4.3 样本平均数差数的分布

$$N_1 = 4,\ n_1 = 2,\ \mu_1 = 3,\ \sigma_1^{\ 2} = 1/2$$
$$N_2 = 3,\ n_2 = 2,\ \mu_2 = 2,\ \sigma_2^{\ 2} = 2/3$$

表4.3 样本平均数差数的分布

$\bar{y}_1 - \bar{y}_2$	f
－1.0	1
－0.5	6
0.0	17
0.5	30
1.0	36
1.5	30
2.0	17
2.5	6
3.0	1
总　和	144

样本平均数差数分布的平均数为：

$$\mu_{\bar{y}_1-\bar{y}_2} = \frac{\sum f(\bar{y}_1 - \bar{y}_2)}{N_1^{n_1} N_2^{n_2}}$$

$$=144/144=1=3-2$$

$$= \mu_1 - \mu_2$$

样本平均数差数分布的方差为：

$$\sigma^2_{\bar{y}_1-\bar{y}_2} = \frac{1}{N_1^{n_1} N_2^{n_2}} \sum f[(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)]^2$$

$$= \frac{7}{12} = \frac{1}{4} + \frac{1}{3} = \frac{1}{2}\Big/2 + \frac{2}{3}\Big/2$$

$$= \frac{\sigma_1^{\ 2}}{n_1} + \frac{\sigma_2^{\ 2}}{n_2} = \sigma^2_{\bar{y}_1} + \sigma^2_{\bar{y}_2}$$

4.3 样本平均数差数的分布

◆ 性质

（1）样本平均数差数的平均数等于总体平均数之差

$$\mu_{\bar{y}_1-\bar{y}_2} = \mu_1 - \mu_2$$

（2）样本平均数差数的方差等于两个样本平均数方差之和

$$\sigma^2_{\bar{y}_1-\bar{y}_2} = \sigma^2_{\bar{y}_1} + \sigma^2_{\bar{y}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

（3）如果 $\bar{y}_1$服从正态分布$N(\mu_1,\sigma^2_{\bar{y}_1})$，$\bar{y}_2$服从正态分布$N(\mu_2,\sigma^2_{\bar{y}_2})$，

则 $(\bar{y}_1-\bar{y}_2)$服从正态分布 $N(\mu_1-\mu_2,\sigma^2_{\bar{y}_1-\bar{y}_2})$。

4.3 样本平均数差数的分布

例题： 已知两个正态总体 $\mu_1-\mu_2=1,\ \sigma^2_{\bar{y}_1}=\frac{1}{4},\ \sigma^2_{\bar{y}_2}=\frac{1}{3}$，

求 $(\bar{y}_1-\bar{y}_2)<0.25$ 的概率。

$$\bar{y}_1-\bar{y}_2 \sim N(\mu_1-\mu_2,\sigma^2_{\bar{y}_1-\bar{y}_2}) \qquad \sigma^2_{\bar{y}_1-\bar{y}_2}=\sigma^2_{\bar{y}_1}+\sigma^2_{\bar{y}_2}$$

$$\sigma_{\bar{y}_1-\bar{y}_2}=\sqrt{\sigma^2_{\bar{y}_1}+\sigma^2_{\bar{y}_2}}=\sqrt{\frac{1}{4}+\frac{1}{3}}=0.7638$$

$$u=\frac{y-\mu}{\sigma} \qquad u=\frac{(\bar{y}_1-\bar{y}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{y}_1-\bar{y}_2}}=(0.25-1)/0.7638=-0.98$$

查附表1：$F(-0.98)=0.1635$

$$P\{(\bar{y}_1-\bar{y}_2)<0.25\}=0.1635$$

第四章 抽样分布

- 4.1 随机抽样与无偏估计
- 4.2 样本平均数的分布
- 4.3 样本平均数差数的分布
- **4.4 t分布**
- 4.5 χ^2分布
- 4.6 F分布

4.4 t 分布

◆ t分布的提出

从正态总体中抽样，其样本平均数$\bar{y}$服从$N(\mu,\sigma_{\bar{y}}^2)$ 。

总体方差已知

$$u=\frac{\bar{y}-\mu}{\sigma_{\bar{y}}}=\frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$$

总体方差一般未知

s^2代替σ^2

$$t=\frac{\bar{y}-\mu}{s_{\bar{y}}}=\frac{\bar{y}-\mu}{\boxed{s/\sqrt{n}}} \qquad (n<30)$$

4.4 *t* 分布

◆ *t*分布的提出

1908年， 英国统计学家W. S. Gosset 于以笔名“student”在论文“The Probable Error of a Mean ”中提出*t*分布，又称为学生氏*t*分布，开创了小样本理论的先河。

THE PROBABLE ERROR OF A MEAN.

By STUDENT.

Introduction.

Any experiment may be regarded as forming an individual of a “population” of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

4.4 t 分布

◆ t分布的概率密度函数：

$$f(t)=\frac{\boxed{\Gamma}(\frac{\boxed{\nu}+1}{2})}{\sqrt{\pi\nu}\,\Gamma(\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-(\frac{\nu+1}{2})},\qquad -\infty<t<+\infty$$

ν:自由度$(n-1)$

Γ: $gamma$函数

◆ t分布的平均数与方差：

$$\mu_t=\int_{-\infty}^{+\infty}f(t)\cdot t\,dt=0$$

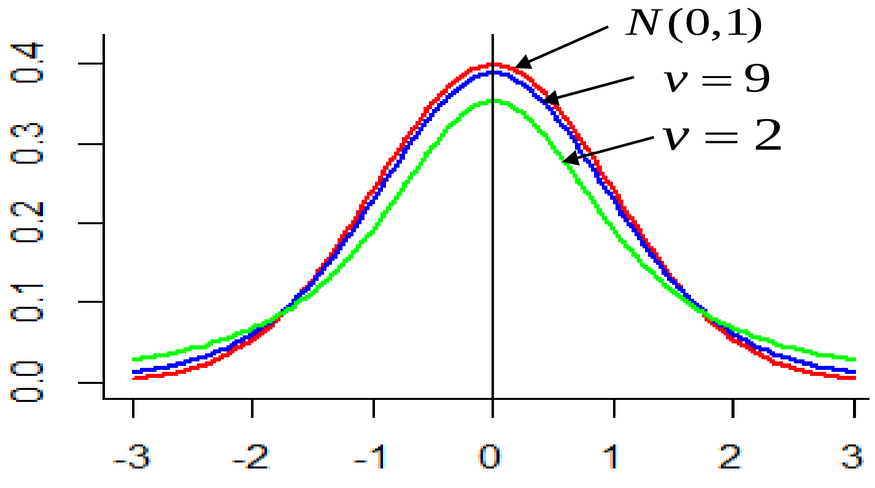
$$\sigma_t^2=\int_{-\infty}^{+\infty}f(t)\cdot(t-\mu_t)^2dt=\frac{\nu}{\nu-2}\quad(\nu>2)$$

4.4 t 分布

◆ 性质

（1）t分布受自由度的制约，每一个自由度都有一条t分布曲线。

（2）t分布密度曲线以t=0为中心，左右对称，且在t＝0时，t分布的概率密度函数取得最大值。

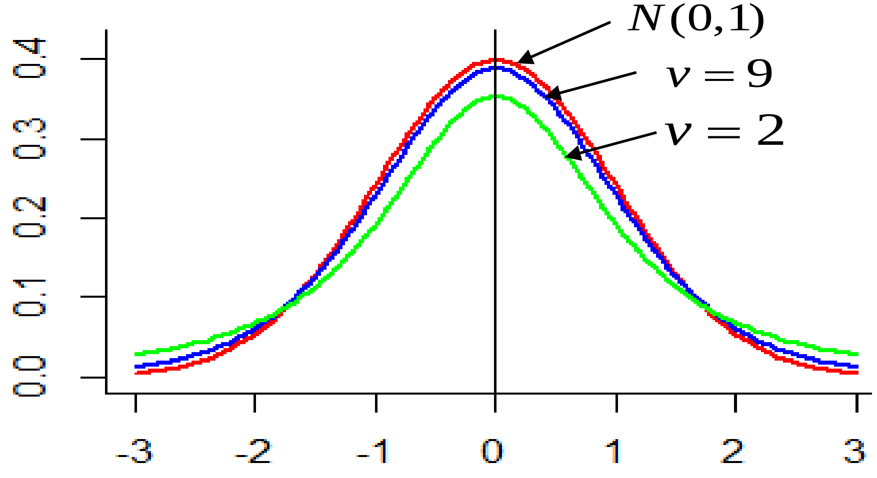


t分布的概率密度曲线

4.4 t 分布

◆ 性质

（3）和标准正态分布相比较，t分布顶端偏低，尾端偏高。ν越大，t分布越趋近于标准正态分布。当$\nu>30$ 时，接近标准正态分布；$\nu\to\infty$ 时，与标准正态分布重合。



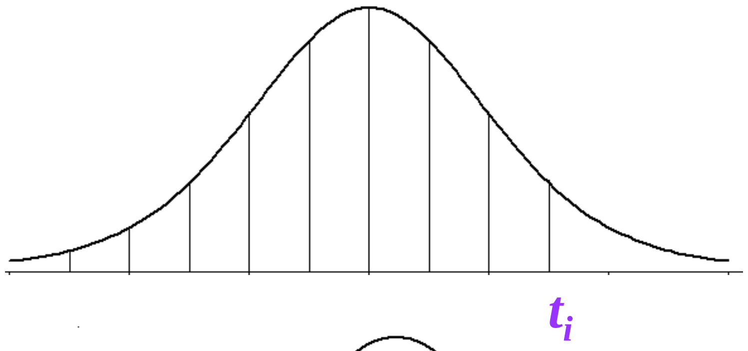
t分布的概率密度曲线

4.4 t 分布

◆ t分布的概率

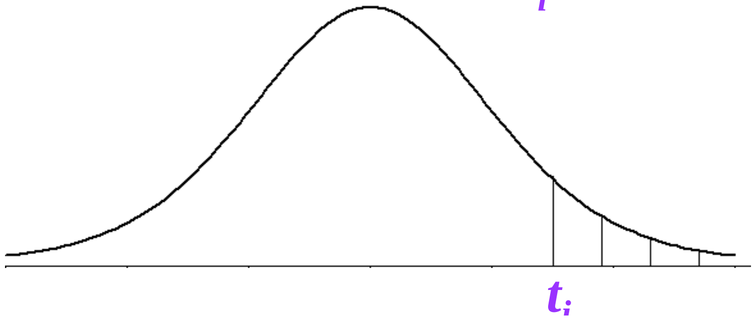
t分布的累积函数$F(t)$：

$$F(t_i) = P\{t \leqslant t_i\} = \int_{-\infty}^{t_i} f(t)dt$$



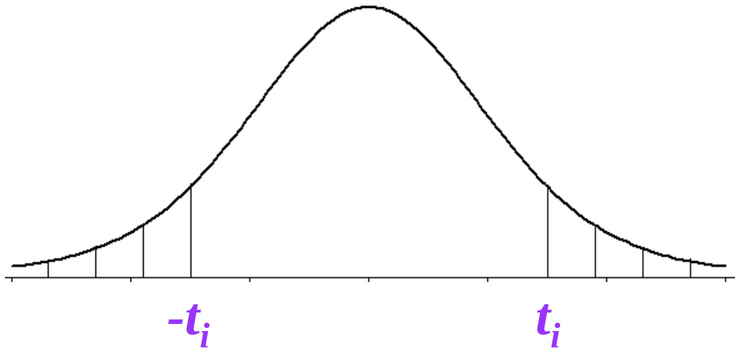
t分布的右尾从t_i到$+\infty$的概率：

$$P\{t > t_i\} = 1 - F(t_i)$$



由于t分布的对称性，故两尾概率为：

$$P\{|t| > t_i\} = 2[1 - F(t_i)]$$



4.4 *t* 分布

◆ *t*分布的概率

附表3列出不同自由度下 *t* 分布的两尾临界值

例：当ν=10时，查询两尾概率等于0.05的临界*t*值

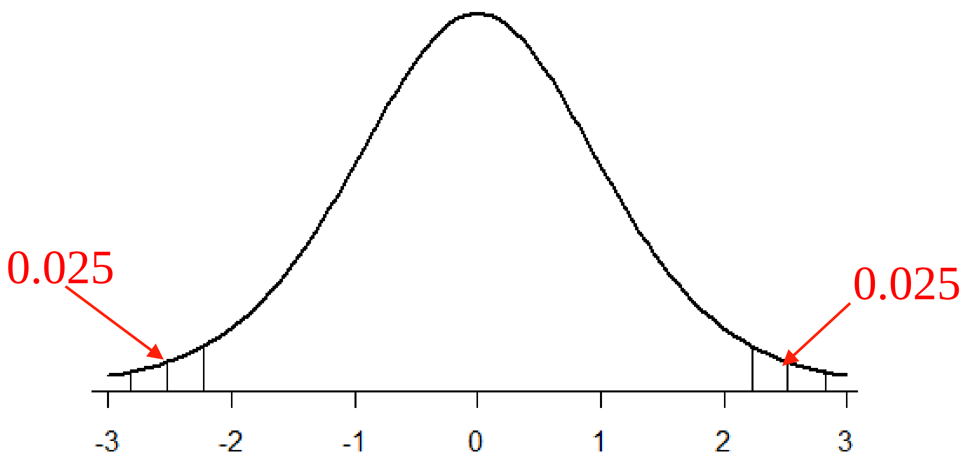
附表 3　t 分布两尾临界值 $t_{\alpha/2,df}$ 表

$$p\{|t|\geq t_{\alpha/2,df}\}=\alpha$$

自由度 *df*	α 0.500	0.200	0.100	0.050	0.025	0.010	0.005
1	1.000	3.078	6.314	12.706	25.452	63.657	127.321
2	0.816	1.886	2.920	4.303	6.205	9.925	14.089
3	0.765	1.638	2.353	3.182	4.177	5.841	7.453
4	0.741	1.533	2.132	2.776	3.495	4.604	5.598
5	0.727	1.476	2.015	2.571	3.163	4.032	4.773
6	0.718	1.440	1.943	2.447	2.969	3.707	4.317
7	0.711	1.415	1.895	2.365	2.841	3.499	4.029
8	0.706	1.397	1.860	2.306	2.752	3.355	3.833
9	0.703	1.383	1.833	2.262	2.685	3.250	3.690
10	0.700	1.372	1.812	2.228	2.634	3.169	3.581

$t_{0.05,10}$=2.228，表示：当ν=10,

$$P\{|t|>2.228\}=P\{t>2.228\}+P\{t<-2.228\}=0.05$$



第四章 抽样分布

扬州大学
YANGZHOU UNIVERSITY

4.5 χ^2分布（卡方分布）

◆ 定义

n个相互独立的标准正态离差u的平方和

$$\chi^2 = u_1^{\ 2} + u_2^{\ 2} + \cdots + u_n^{\ 2} = \sum_{i=1}^{n} u_i^{\ 2} = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{\sigma^2} \sim \chi^2_{v=n}$$

样本：

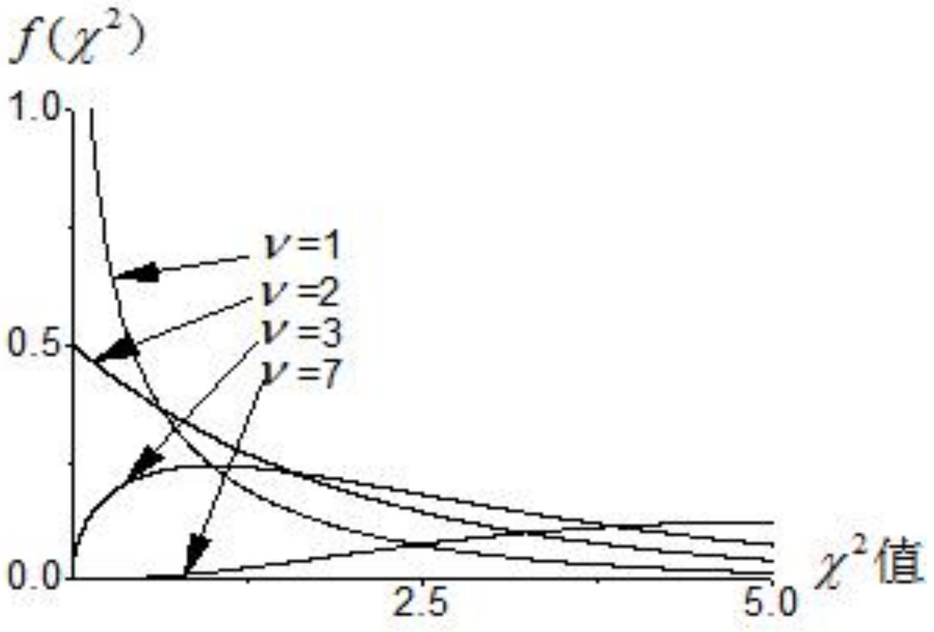
$$s^2 = \frac{\sum(y - \bar{y})^2}{n-1}$$

$$\chi^2 = \frac{\sum_{i=1}^{n}(y_i - \boxed{\bar{y}})^2}{\sigma^2} = \frac{\boxed{(n-1)s^2}}{\sigma^2} \sim \chi^2_{v=n\text{-}1}$$

4.5 χ^2分布

◆ χ^2分布的概率密度函数

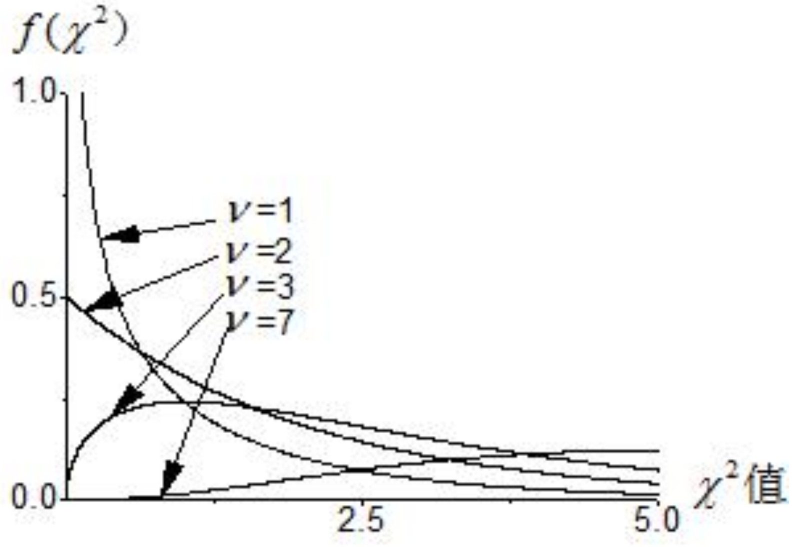
$$f(\chi^2)=\frac{(\chi^2)^{\frac{v}{2}-1}e^{-\frac{1}{2}\chi^2}}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})}$$



不同自由度下χ^2分布的概率密度曲线

4.5 χ^2 分布

◆ χ^2分布的性质



① 定义域 $\chi^2 \in [0, +\infty)$

② 平均数 $\mu_{\chi^2} = \nu$

③ 在$v = 1$时，分布极度左偏；在$v = 2$ 时，曲线在$\chi^2 = 0$处最高；随着v增大，曲线渐趋左右对称。

④ 当$v > 30$，χ^2分布趋于正态分布 $\sqrt{2\chi^2} \sim N(\sqrt{2\nu - 1}, 1)$

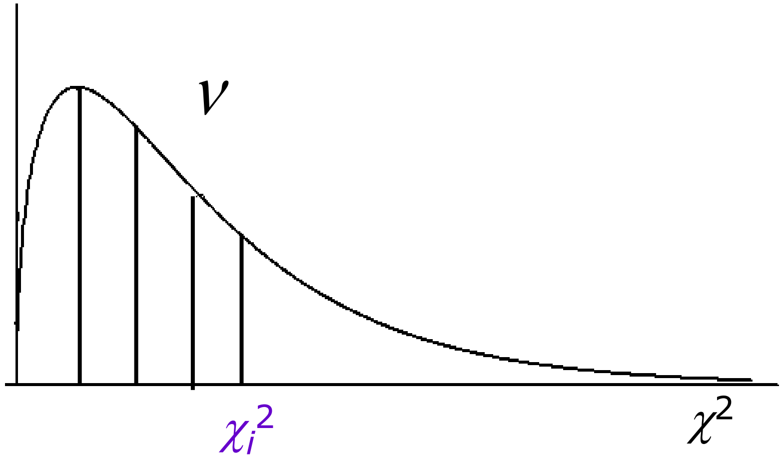
$$u = \sqrt{2\chi^2} - \sqrt{2\nu - 1}$$

4.5 χ^2 分布

◆ χ^2分布的概率计算

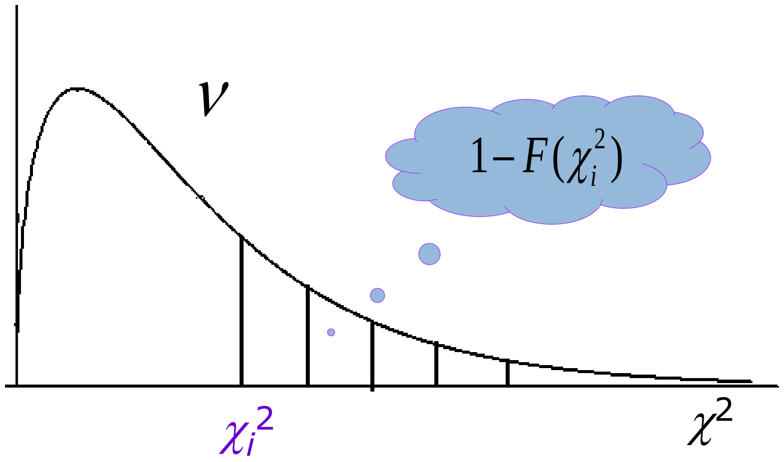
χ^2分布的累积函数$F(\chi^2)$：

$$F(\chi_i^2)=P\{\chi^2\leqslant\chi_i^2\}=\int_0^{\chi_i^2} f(\chi^2)d(\chi^2)$$



χ^2分布的右尾从χ_i^2到$+\infty$的概率：

$$P\{\chi^2>\chi_i^2\}=1-F(\chi_i^2)=\int_{\chi_i^2}^{+\infty} f(\chi^2)d(\chi^2)$$



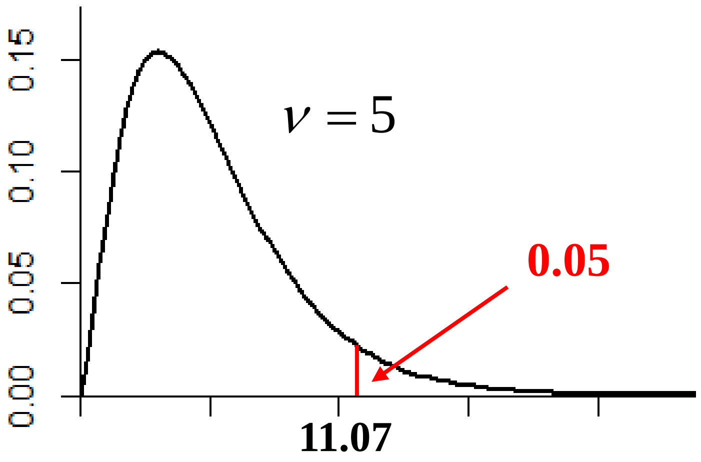
4.5 χ^2 分布

◆ χ^2分布的概率计算

例：当ν=5时，查询右尾概率等于0.05的临界χ^2值

$$\chi^2_{0.05,5}=11.07$$

$$P(\chi^2>11.07)=0.05$$



附表7 χ^2分布右尾临界值 $\chi^2_{\alpha,df}$表

$$P\{\chi^2\geqslant\chi^2_{\alpha,df}\}=\alpha$$

自由度 df	α									
	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1					0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.69	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19

例题： 已知某总体$\sigma^2=10$，试求当随机抽取$n=40$的样本时，求$P\{s^2\leqslant 9\}=?$

由于$n>30$，可作正态近似

$$\chi^2=\frac{(n-1)s^2}{\sigma^2}=\frac{(40-1)\times 9}{10}=35.1$$

$$u=\sqrt{2\chi^2}-\sqrt{2v-1}$$

$$=\sqrt{2\times 35.1}-\sqrt{2\times 39-1}=-0.40$$

查附表1：$F(-0.40)=0.3446$

$$P\{s^2\leqslant 9\}=0.3446$$

第四章 抽样分布

4.6 *F*分布

◆ 定义

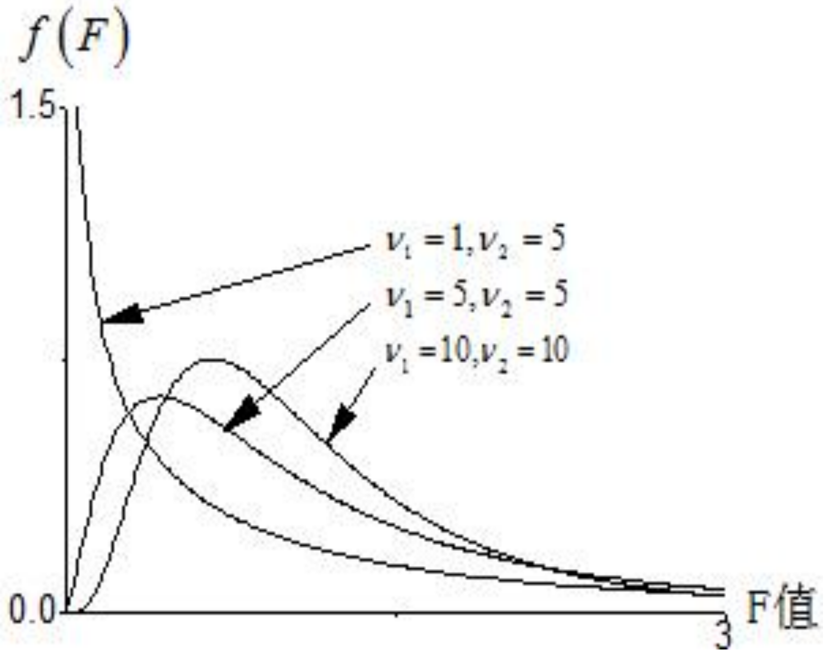
在一正态总体N（μ，σ^2）中随机抽取样本容量为n_1和n_2的两个样本，其样本方差（均方）的比值定义为F

$$F = \frac{s_1^{\ 2}}{s_2^{\ 2}} \sim F_{\nu_1,\nu_2}\text{分布} \qquad \nu_1 = n_1 - 1, \nu_2 = n_2 - 1$$

4.6 F分布

◆ F分布的概率密度函数

$$f(F)=\frac{\Gamma(\frac{v_1+v_2}{2})}{\Gamma(\frac{v_1}{2})\Gamma(\frac{v_2}{2})}v_1^{\frac{v_1}{2}}v_2^{\frac{v_2}{2}}\frac{F^{\frac{v_1}{2}-1}}{(v_1F+v_2)^{\frac{v_1+v_2}{2}}}$$



◆ F分布的性质

（1）F分布的取值区间为$[0,+\infty)$

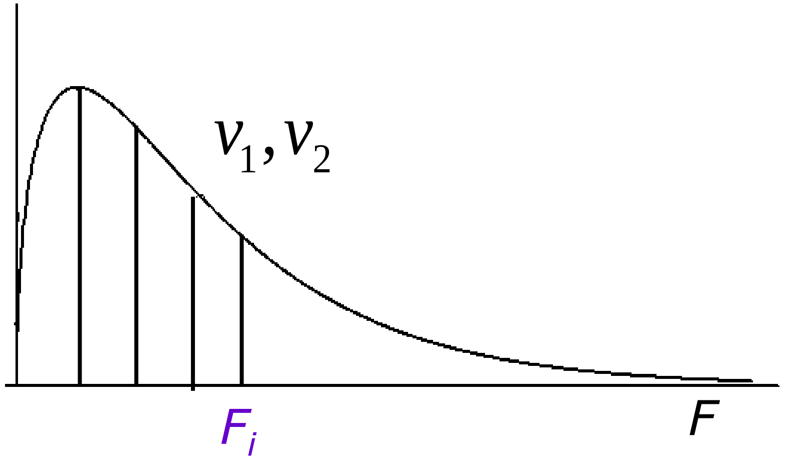
（2）F分布曲线的形状仅决定于v_1和v_2

4.6 F分布

◆ F分布的概率计算

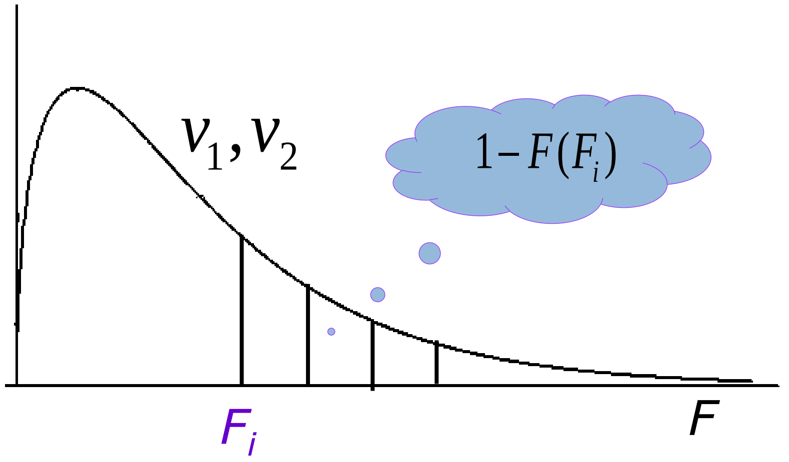
F分布的累积函数$F(F_i)$：

$$F(F_i)=P\{F\leqslant F_i\}=\int_0^{F_i} f(F)dF$$



F分布右尾从F_i到$+\infty$的概率：

$$P\{F>F_i\}=1-F(F_i)$$



4.6 F分布

- F分布的概率计算

附表 4　F 分布右尾临界值 $F_{\alpha,\ df_1,\ df_2}$ 表

df_2	df_1													
	1	2	3	4	5	6	7	8	9	10	11	12	14	16
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246
	4 052	5 000	5 403	5 625	5 764	5 859	5 928	5 981	6 022	6 056	6 082	6 106	6 143	6 170
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.43
	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.43	99.44
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.71	8.69
	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.13	27.05	26.92	26.83
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.87	5.84
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.25	14.15
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.64	4.60
	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.77	9.68

$$P\left\{F > F_{\alpha,v_1,v_2}\right\} = 0.05（上）$$

$$P\left\{F > F_{\alpha,v_1,v_2}\right\} = 0.01（下）$$

当$\boldsymbol{v_1}$ =1，　$\boldsymbol{v_2}$ =5时，查询右尾概率等于0.05和0.01的临界F值

$F_{0.05,1,5}$=6.61，$F_{0.01,1,5}$=16.26

$$P\left\{F > 6.61\right\} = 0.05$$

$$P\left\{F > 16.26\right\} = 0.01$$