

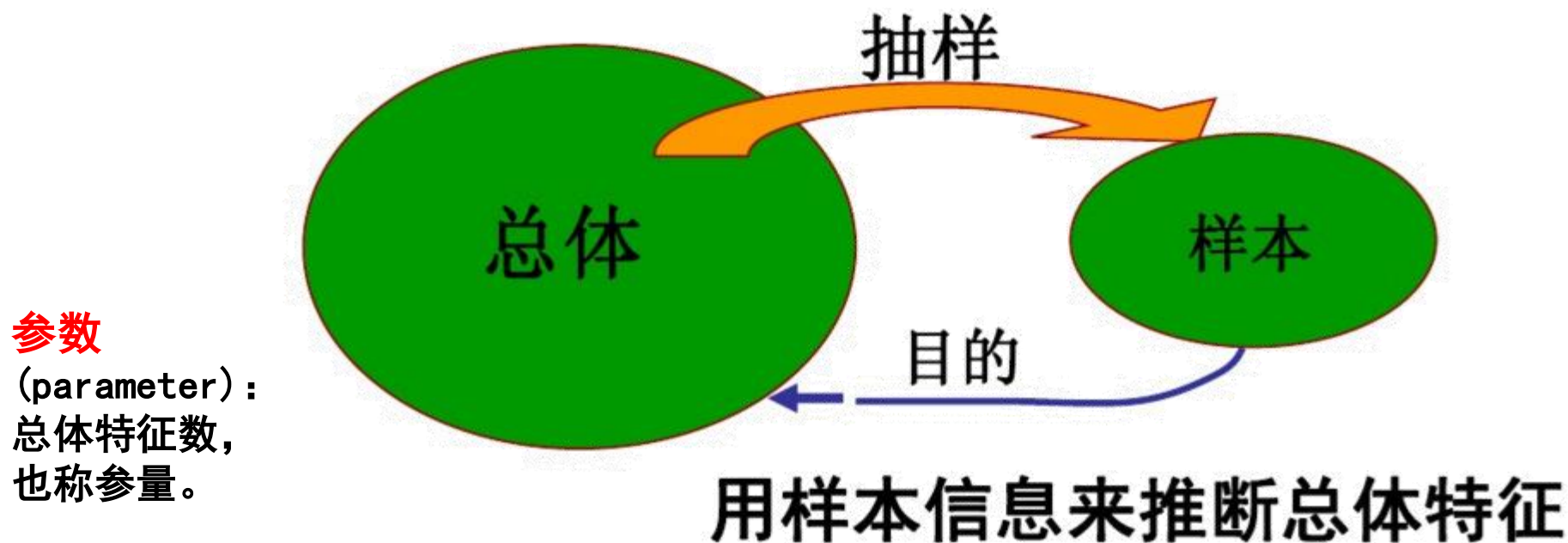
第3章

统计量分布-抽样分布



统计学基本问题：

研究总体与样本的关系



参数
(parameter):
总体特征数,
也称参量。

统计量 (statistic):
样本特征数, 也称
统计数, 由样本算
出的量, 或者说统
计量就是样本的函
数。统计量只依赖
于样本, 而不能与
任何未知的量有关。

抽样分布: 样本统
计量的分布

样本统计量推断总体参数, 以抽样分布为基础
统计推断: 假设检验, 参数估计

生物统计的最基本的问题是研究 总体与样本的关系

总体与样本的关系可以从两个方面研究

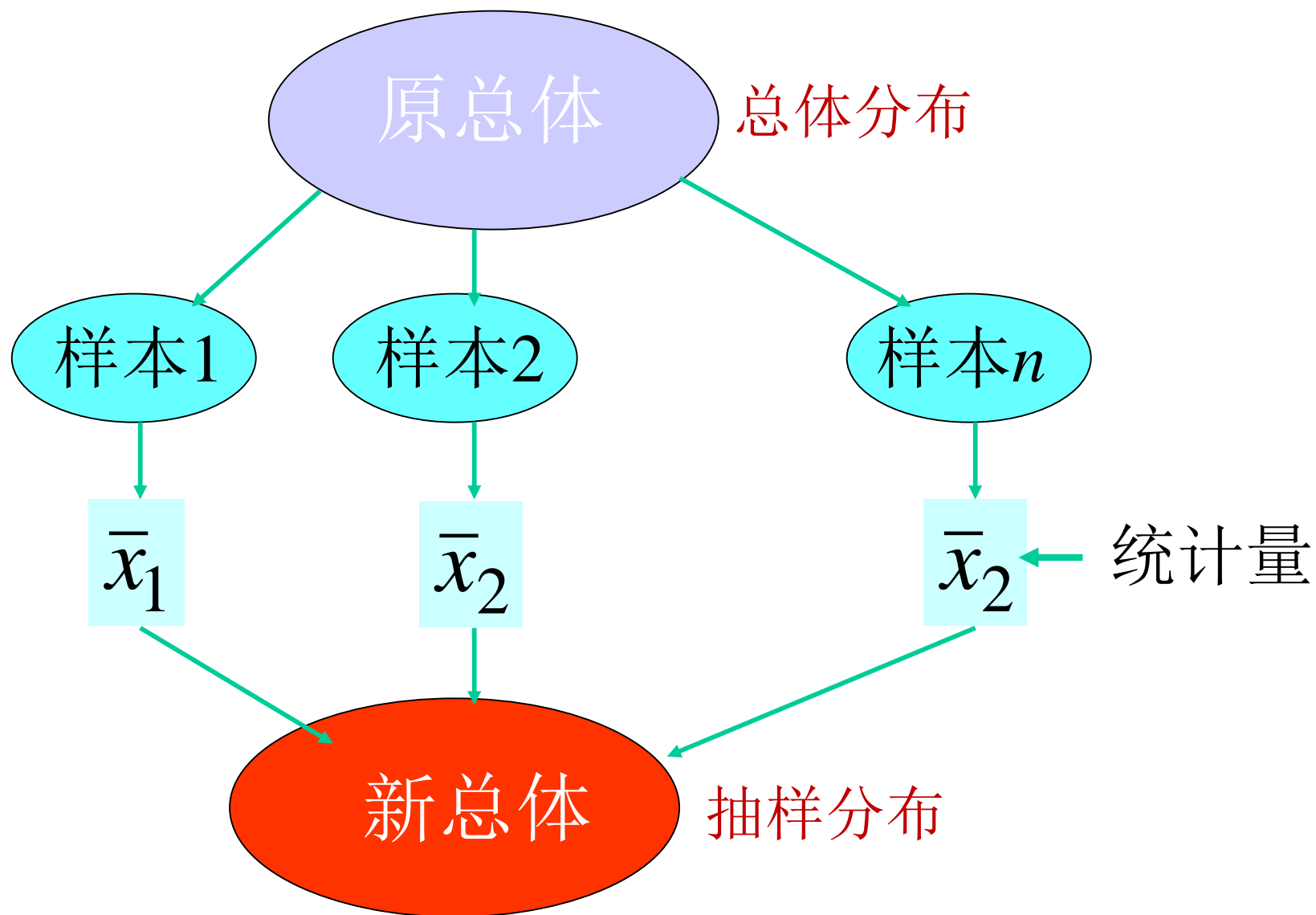
- ① 由已知的总体，研究样本的分布规律，即由总体到样本的过程；研究样本的各种统计量的概率分布，即所谓的**抽样分布(sampling distribution)**
- ② 由样本推断未知的总体，属于从样本到总体的研究过程。即**统计推断问题**。

统计量分布-抽样分布(sampling distribution)

- 样本是随机变量, 有一定的概率分布, 即样本分布. 统计量是样本的已知函数, 则它也有其概率分布, 且这个概率分布在原则上可由样本分布算出. **统计量的概率分布称为 (该统计量的) 抽样分布.**

抽样分布的概念

就比如说调查一所中学的所有学生的身高,这就构成了总体,从中随机抽取300个人,这300个人就组成一个样本分布.之后再抽取若干个300人组成的样本,每个样本的平均数的分布就是抽样分布。



样本统计量

- 在抽样估计中，用来反映样本总体数量特征的指标称为样本指标，也称为样本统计量或估计量，是根据样本资料计算的、用以估计或推断相应总体指标的综合指标。

常见的样本统计量有：

变量总体	属性总体
样本平均数 \bar{x}	样本比例（样本成数） p
样本标准差 s 或方差 s^2	样本比例标准差 s_p 或方差 s_p^2

抽样分布的概念

- **样本统计量**的概率分布称为抽样分布(sampling distribution)

根据样本对总体做出估计和推断，并不是直接用样本本身，而是用样本的统计量来对总体做出估计和判断。但由于从总体中抽取的样本提供的信息仅是总体的一部分，因此它不能提供完全准确的信息，必然存在着一定的误差。即，对于**样本容量相同的多次随机抽样**，得到样本函数的观察值也是不同的，且其取值有一定的概率，即**统计量也是一个随机变量**，因而也有它的分布，称为抽样分布(sampling distribution)。

相关概念

- ◆ 总体分布：所有元素出现概率的分布.是简单意义上的随机变量对应的频次分布.
- ◆ 样本分布：样本分布有区别于总体分布,它是从总体中按一定的分组标志选出来的部分样本容量.选择的样本在随机变量上的对应的频次分布,样本分布实际上也在趋向总体分布.
- ◆ 抽样分布：是对**样本统计量**概率分布的一种描述方式. 抽样分布是一种概率分布,随机变量是样本统计量.

抽样分布的重要性

- 统计推断的结果取决于抽得的样本，而样本受随机性的干扰，因而推断的结果也是随机的：一个整体上看来较好的推断方法，在个别情况下可以给出不好的结果. 反之亦然. 因此，统计推断方法优良的指标只能是整体性的，即取决于所用统计量的抽样分布. 总之，要想得到一种特定的统计推断方法的全面了解，必须确定其抽样分布。

基于正态总体的抽样分布

- 我们所研究的抽样分布，全部都是建立在正态分布基础上的。或者说全部都是从正态总体中进行抽样的。但是根据中心极限定理，从一个非正态分布的总体中抽取的容量为 n 的样本，当 n 充分大时，样本平均数渐近服从正态分布。因此平均数的抽样分布对正态性的要求并不是十分严格，但方差的抽样分布，对总体的正态性的要求是十分严格的。

第3章

统计量的分布

3.1 样本平均数的分布





3.1 样本平均数的分布

例：样本平均数分布

现有一 $N=3$ 的近似正态总体，具有变量3,4,5，可以求出

$$\mu = 4, \sigma^2 = 0.6667, \sigma = 0.8165。$$

现以 $n=2$ 作独立的有放回式抽样。

总共可得到 $N^n = 3^2 = 9$ 个样本



抽样分布

样本编号	样本值	\bar{x}	s^2	s
1	3,3	3.0	0.0	0.0000
2	3,4	3.5	0.5	0.7071
3	3,5	4.0	2.0	1.4142
4	4,3	3.5	0.5	0.7071
5	4,4	4.0	0.0	0.0000
6	4,5	4.5	0.5	0.7071
7	5,3	4.0	2.0	1.4142
8	5,4	4.5	0.5	0.7071
9	5,5	5.0	0.0	0.0000
Σ		36.0	6.0	5.6568
平均		4.0	0.6667	0.6258

$\mu = 4$

$\sigma^2 = 0.6667$

$\sigma = 0.8165$

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

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

$\mu_s \neq \sigma$

由于从总体中抽出的样本为每一个可能样本，且每个样本中的变量均为随机变量，所以其样本平均数也为随机变量，也形成一定的理论分布，这种理论分布称为**样本平均数的概率分布**，或称**样本平均数的分布**。

样本平均数的平均数：

$$\mu_{\bar{x}}$$

样本平均数的方差：

$$\sigma_{\bar{x}}^2$$



n=2

对N=3(3,4,5), n=2抽样试验所得的9个样本平均数, 整理成次数分布表。

\bar{x}	f	f \bar{x}	f \bar{x}^2
3.0	1	3	9.0
3.5	2	7	24.5
4.0	3	12	48.0
4.5	2	9	40.5
5.0	1	5	25.0
Σ	9	36	147.0

n=2

\bar{x}	f	$f\bar{x}$	$f\bar{x}^2$
3.0	1	3	9.0
3.5	2	7	24.5
4.0	3	12	48.0
4.5	2	9	40.5
5.0	1	5	25.0
Σ	9	36	147.0

3,4,5
 $\mu = 4$
 $\sigma^2 = 0.6667$

$$\mu_{\bar{x}} = \frac{\sum f \bar{x}}{N^n} = \frac{36}{9} = 4$$

$$= \mu$$

$$\sigma_{\bar{x}}^2 = \frac{1}{N^n} \left(\sum f \bar{x}^2 - \frac{(\sum f \bar{x})^2}{N^n} \right) = 0.3333$$

$$= \frac{\sigma^2}{n}$$

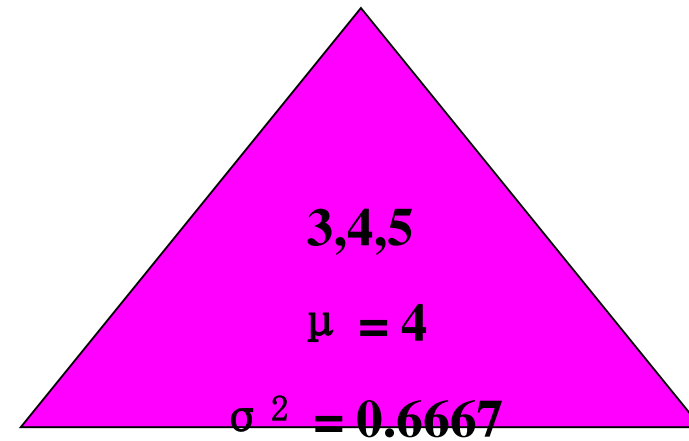
n=4

如果对这个
N=3(3,4,5) 所组成的
总体，再进行n=4的
抽样试验，则可得81
个样本平均数，将其
整理成次数分布表。

\bar{x}	f	$f\bar{x}$	$f\bar{x}^2$
3.00	1	3	9.00
3.25	4	13	42.25
3.50	10	35	122.50
3.75	16	60	225.00
4.00	19	76	304.00
4.25	16	68	289.00
4.50	10	45	202.50
4.75	4	19	90.25
5.00	1	5	25.00
Σ	81	324	1309.50

n=4

\bar{x}	f	$f\bar{x}$	$f\bar{x}^2$
3.00	1	3	9.00
3.25	4	13	42.25
3.50	10	35	122.50
3.75	16	60	225.00
4.00	19	76	304.00
4.25	16	68	289.00
4.50	10	45	202.50
4.75	4	19	90.25
5.00	1	5	25.00
Σ	81	324	1309.50



$$\mu_{\bar{x}} = \frac{\sum f \bar{x}}{N^n} = \frac{324}{81} = 4$$

$$= \mu$$

$$\sigma_{\bar{x}}^2 = \frac{1}{N^n} \left(\sum f \bar{x}^2 - \frac{(\sum f \bar{x})^2}{N^n} \right) = 0.1667$$

$$= \frac{\sigma^2}{n}$$

样本平均数的分布

- 生物学中遇到的总体都是很大的，几乎都是无限的。从这样的样本中抽取含量为 n 的样本，样本统计量的个数也是无限的，不可能用上述方法做抽样研究。但可以通过收集“足够多的”样本值，来研究样本统计量的分布，常常称为“Monte Carlo”方法就可以由样本来推断总体了
- 对于正态总体，可以用数学方法推导出样本统计量的分布。以下讲几种从正态总体中抽取的样本统计量的分布。有了严格的样本分布，就可以由样本来推断总体了。

基于正态总体的抽样分布

- 我们所研究的抽样分布，全部都是建立在正态分布基础上的。或者说全部都是人从正态总体中进行抽样的。但是根据中心极限定理，从一个非正态分布的总体中抽取的容量为 n 的样本，当 n 充分大时，样本平均数渐近服从正态分布。因此平均数的抽样分布对正态性的要求并不是十分严格，但方差的抽样分布，对总体的正态性的要求是十分严格的。

样本平均数分布的基本性质

(1) 样本平均数分布的平均数=总体平均数。

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

(2) 样本平均数分布的方差=总体方差除以样本容量。

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

Some Important Characteristics of the Sampling Distribution of \bar{X}

Mean and Variance of \bar{X}

If X_1, X_2, \dots, X_n is a random sample of size n from any distribution with mean μ and variance σ^2 , then:

1. The mean of \bar{X} is

$$\mu_{\bar{X}} = \mu.$$

2. The variance of \bar{X} is

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

3. The standard deviation of \bar{X} is called the standard error of \bar{X} and is defined by

$$\sigma_{\bar{X}} = \sqrt{\sigma_{\bar{X}}^2} = \frac{\sigma}{\sqrt{n}}.$$

It may be noted here that the square root of the variance of a statistic is termed as standard error. In this case, the standard error of the sample mean is $\frac{\sigma}{\sqrt{n}}$.

样本平均数的标准误差（标准误） (standard error of mean)

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

标准误反映了样本平均数 \bar{x} 的抽样误差，即精确性的高低。

标准误大，各样本平均数间差异程度大，样本平均数的精确性低。

标准误小，各样本平均数间差异程度小，样本平均数的精确性高。

标准误的大小与原总体的标准差 σ 成正比，与样本含量 n 的平方根成反比。

从某特定总体抽样，因为 σ 是一定值，所以只有增大样本容量，才能降低样本平均数的抽样误差。



3.1.1 从一个正态总体中抽取的样本统计量的分布



标准差已知时平均数的分布

1. 抽自正态分布总体

如果从**正态分布总体** $N(\mu, \sigma^2)$ 进行独立随机抽样含量为 n 的样本 X_1, X_2, \dots, X_n , 则样本平均数

$\bar{Y} = \sum_{i=1}^n X_i$ 服从均值为 μ , 方差为 σ^2/n 的正态分布, 记作:

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

将平均数 \bar{Y} 标准化，则

$$u = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}}$$
$$u \sim N(0, 1)$$

其中 σ/\sqrt{n} 正态总体中进行抽样为平均数的标准误差
(standard error of mean)

2. 抽自非正态分布总体

中心极限定理(central limit theorem)

如果被抽总体不是正态分布总体，但具有平均数 μ 和方差 σ^2 ，随样本容量 n 的不断增大，样本平均数 \bar{Y} 的分布也越来越接近正态分布，且具有平均数 μ ，方差 σ^2 / n

不论总体为何种分布，只要是大样本，就可运用中心极限定理，认为样本平均数的分布是正态分布，在计算样本平均数出现的概率时，样本平均数可按下式进行标准化。

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

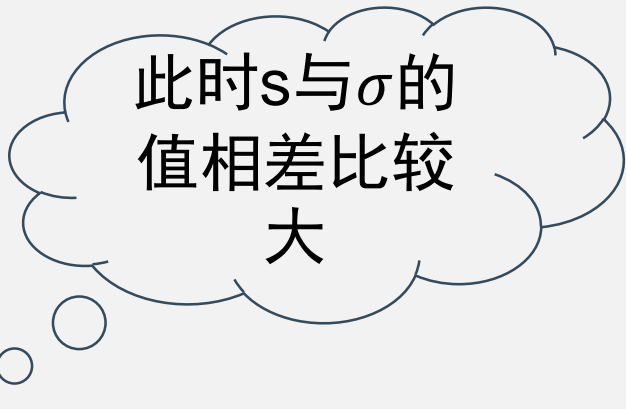


标准差未知时的平均数的分布

标准差未知时的样本平均数的分布---t 分布

若总体的方差是未知的，即标准差 σ 未知，可以用样本的标准差 s 代替总体的标准差 σ ，

则变量 $u = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ 变为 $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$



此时 s 与 σ 的值相差比较大

u 符合 $N(0, 1)$ 分布， t 则不服从标准正态分布，而是服从具有 $(n-1)$ 自由度的 t 分布，其中 $\frac{s}{\sqrt{n}}$

称为**样本标准误差**。

标准差未知时的平均数分布-t分布(t-distribution)

- 总体标准差 σ 未知，可用样本标准差 s 估计总体标准差 σ 标准化变量
$$u = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s / \sqrt{n}} \neq N(\mu, \sigma^2)$$
- 小样本 ($n < 30$)
- 而是服从自由度为 $n-1$ 的 **t分布**。
- 自由度指独立观测值的个数，因为计算 s 时所使用的 n 个观测值受平均数 \bar{x} 的约束，就等于有一个观测值不能独立取值。 **t分布同样要求总体是正态的。**

$$x \rightarrow N(\mu, \sigma^2)$$

t分布

当 σ^2 已知

$$u = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \rightarrow N(0,1)$$

当 σ^2 未知,
且 $n > 30$

$$u = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s / \sqrt{n}} \rightarrow N(0,1)$$

当 σ^2 未知,
且 $n < 30$

$$u = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s / \sqrt{n}} \neq N(\mu, \sigma^2)$$

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} =$$

t分布是英国统计学家Gosset 1908年以笔名“student”所发表的论文提出的，因此又称为学生氏t分布。

t分布概率密度函数

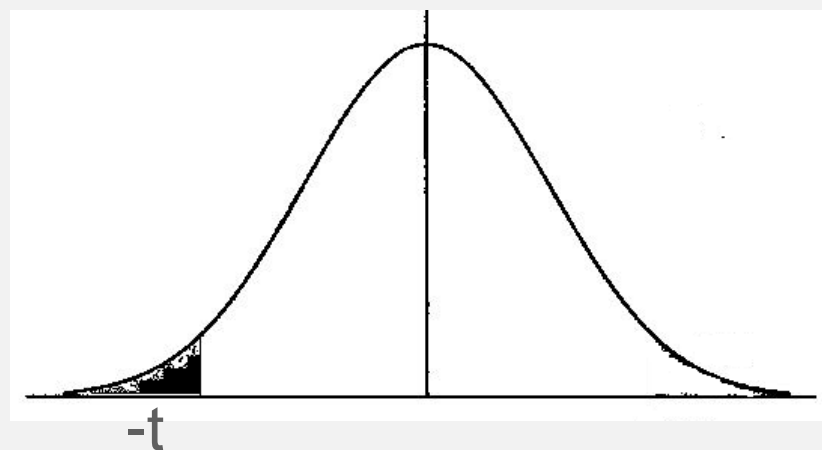
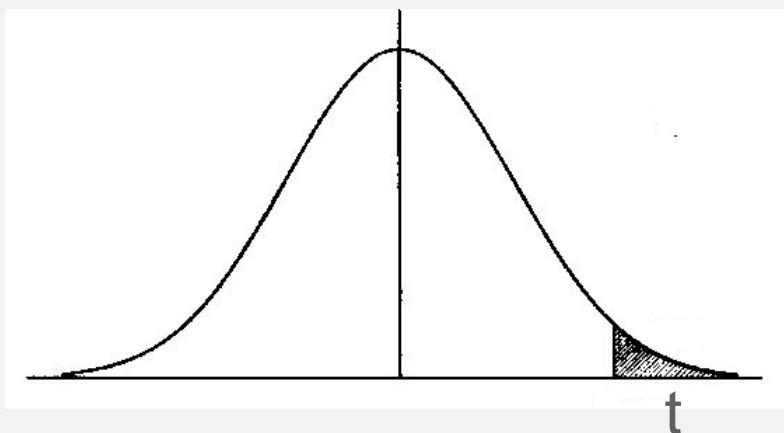
$$f(t) = \frac{\Gamma(\frac{df+1}{2})}{\sqrt{\pi df} \Gamma(\frac{df}{2})} \left(1 + \frac{t^2}{df}\right)^{-\frac{df+1}{2}}$$

t 分布的分位数：

单侧分位数 $\left\{ \begin{array}{l} \text{上侧分位数} \\ \text{下侧分位数} \end{array} \right.$

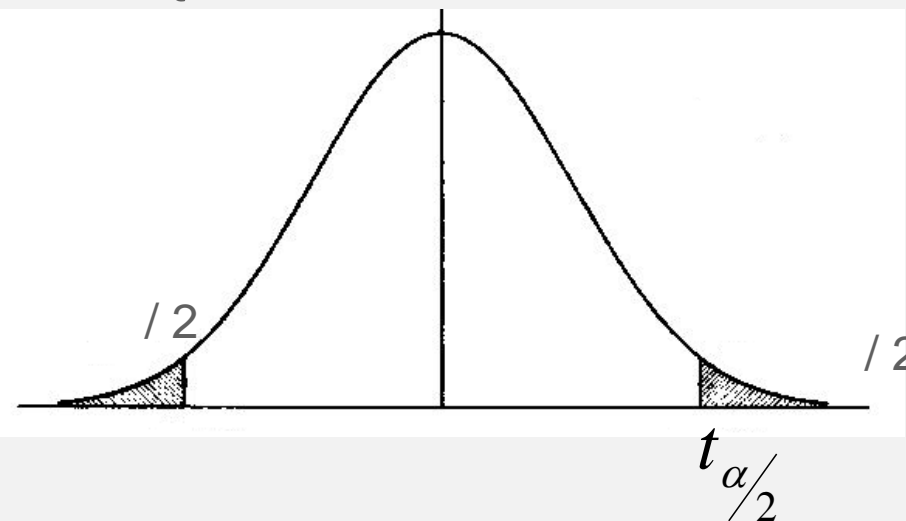
$$P(t \geq t_{\alpha}) = \alpha \quad t_{\alpha}$$

$$P(t \leq -t_{\alpha}) = \alpha \quad -t_{\alpha}$$

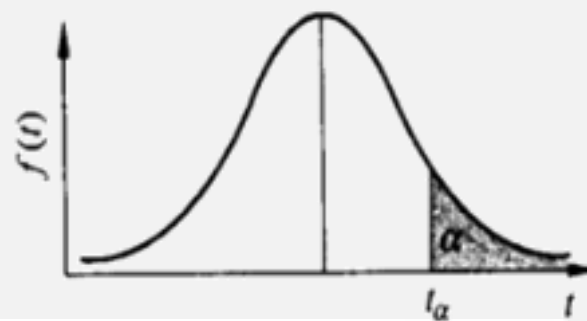


双侧分位数： $P(|t| \geq t_{\alpha/2}) = \alpha$

$t_{\frac{\alpha}{2}}$ 或者 $t_{\alpha} \text{ (双侧)}$



t分布的临界值表



单侧	$\alpha=0.10$	0.05	0.025	0.01	0.005
双侧	$\alpha=0.20$	0.10	0.05	0.02	0.01
$V=1$	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012

t分布的平均数 μ_t
和方差 σ_t^2

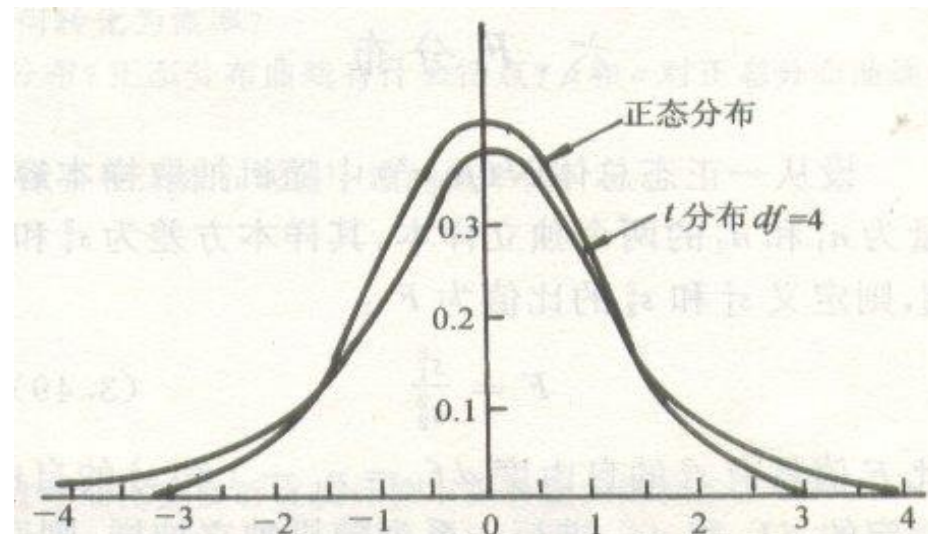


图 3.12 正态分布曲线与 t 分布曲线的比较

$$\mu_t = 0 (df > 1)$$

$$\sigma_t^2 = \frac{df}{df - 2} (df > 2)$$

特征

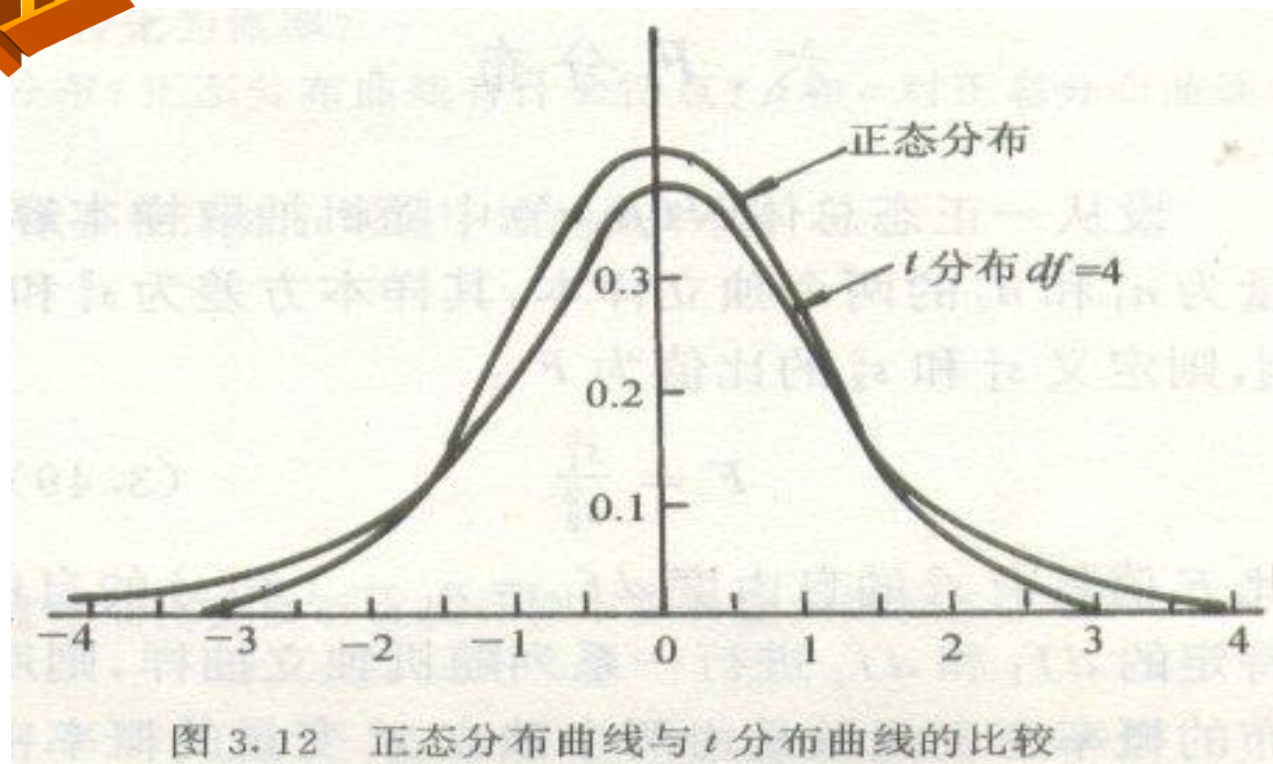
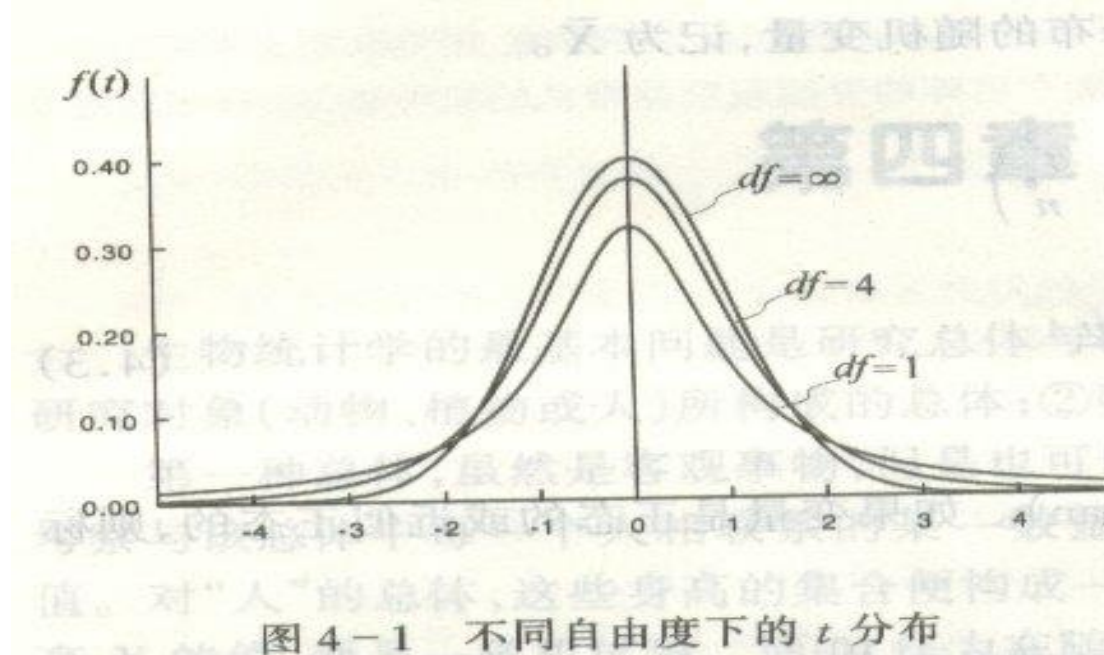


图 3.12 正态分布曲线与 t 分布曲线的比较

(1) t 分布曲线是左右对称的，围绕平均数 $\mu_t=0$ 向两侧递降。

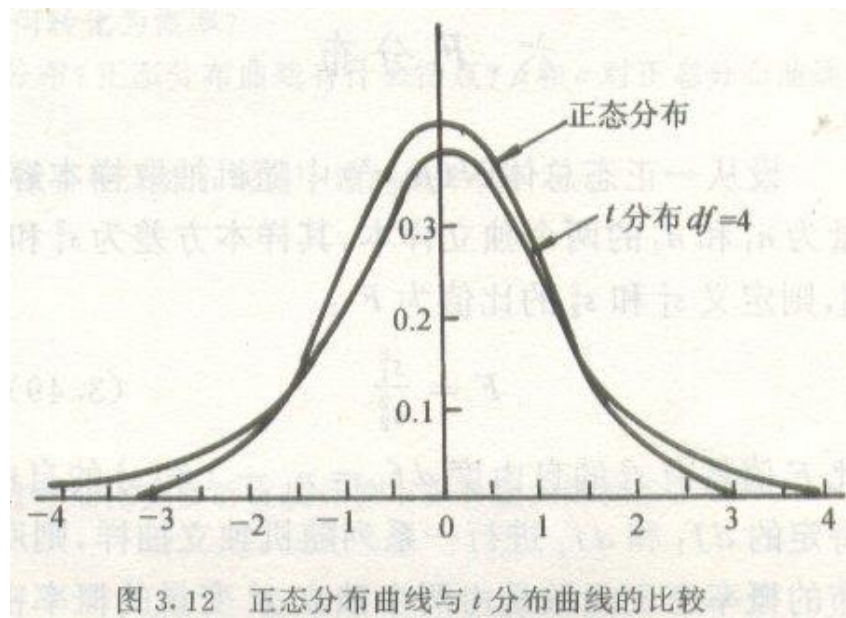
特征

对于不同的自由度， t 分布有不同的曲线。



(2) t 分布受自由度 $df=n-1$ 的制约，每个自由度都有一条 t 分布曲线。

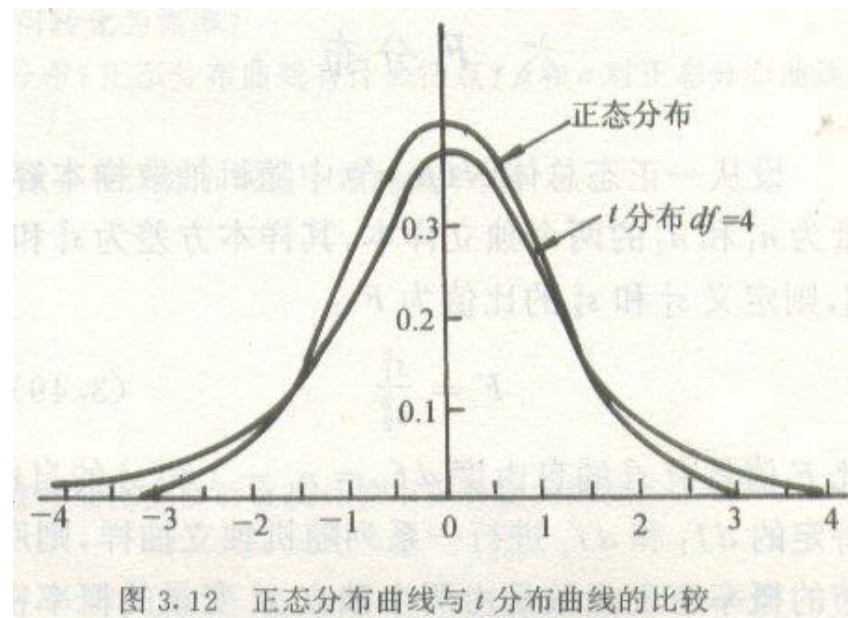
特征



(3) 和正态分布相比， t 分布顶端偏低，尾部偏高，自由度 $df > 30$ 时，其曲线接近正态分布曲线， $df \rightarrow \infty$ 时则和正态分布曲线重合。

$$S_{\bar{x}}$$

$$S_{\bar{x}} \Rightarrow \sigma_{\bar{x}}$$



t 分布曲线与横轴所围成的面积为1。

同标准正态分布曲线一样，统计应用中最为关心的是 t 分布曲线下的面积（即概率 P ）与横轴 t 值间关系。

为使用方便，统计学家编制不同自由度 df 下的 t 值表。

附表 3 t 值表(双尾)

自由度 (df)	概 率 值 (P)								
	0.500	0.400	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	1.000	1.376	3.078	6.314	12.706	25.452	63.657		
2	0.816	1.061	1.886	2.920	4.303	6.205	9.925	14.089	31.598
3	0.765	0.978	1.638	2.353	3.182	4.176	5.841	7.453	12.941
4	0.741	0.941	1.533	2.132	2.776	3.495	4.604	5.598	8.610
5	0.727	0.920	1.476	2.015	2.571	3.163	4.032	4.773	6.859
6	0.718	0.906	1.440	1.943	2.447	2.969	3.707	4.317	5.959
7	0.711	0.896	1.415	1.895	2.365	2.841	3.499	4.029	5.405
8	0.706	0.889	1.397	1.860	2.306	2.752	3.355	3.832	5.041
9	0.703	0.883	1.383	1.833	2.262	2.685	3.250	3.690	4.781
10	0.700	0.879	1.372	1.812	2.228	2.634	3.169	3.581	4.587
11	0.697	0.876	1.363	1.796	2.201	2.593	3.106	3.497	4.437
12	0.695	0.873	1.356	1.782	2.179	2.560	3.056	3.428	4.318
13	0.694	0.870	1.350	1.771	2.160	2.533	3.012	3.372	4.221
14	0.692	0.868	1.345	1.761	2.145	2.510	2.977	3.326	4.140
15	0.691	0.866	1.341	1.753	2.131	2.490	2.947	3.286	4.073
16	0.690	0.865	1.337	1.746	2.120	2.473	2.921	3.252	4.015
17	0.689	0.863	1.333	1.740	2.110	2.458	2.898	3.222	3.965
18	0.688	0.862	1.330	1.734	2.101	2.445	2.878	3.197	3.922
19	0.688	0.861	1.328	1.729	2.093	2.433	2.861	3.174	3.883
20	0.687	0.860	1.325	1.725	2.086	2.423	2.845	3.153	3.850
21	0.686	0.859	1.323	1.721	2.080	2.414	2.831	3.135	3.819
22	0.686	0.858	1.321	1.717	2.074	2.406	2.819	3.119	3.792
23	0.685	0.858	1.319	1.714	2.069	2.398	2.807	3.104	3.767
24	0.685	0.857	1.318	1.711	2.064	2.391	2.797	3.090	3.745
25	0.684	0.856	1.316	1.708	2.060	2.385	2.787	3.078	3.725
26	0.684	0.856	1.315	1.706	2.056	2.379	2.779	3.067	3.707
27	0.684	0.855	1.314	1.703	2.052	2.373	2.771	3.056	3.690
28	0.683	0.855	1.313	1.701	2.048	2.368	2.763	3.047	3.674
29	0.683	0.854	1.311	1.699	2.045	2.364	2.756	3.038	3.659
30	0.683	0.854	1.310	1.697	2.042	2.360	2.750	3.030	3.646
40	0.681	0.851	1.303	1.684	2.021	2.329	2.704	2.971	3.551
60	0.679	0.848	1.296	1.671	2.000	2.299	2.660	2.915	3.460
80	0.678	0.847	1.293	1.665	1.989	2.284	2.638	2.887	3.415
120	0.677	0.845	1.289	1.658	1.980	2.270	2.617	2.860	3.373
∞	0.675	0.842	1.282	1.645	1.960	2.241	2.576	2.807	3.291

- $P\{t(df) > t_\alpha\} = \alpha$

1 在相同的自由度 df 时， t 值越大，概率 P 越小。

2 在相同 t 值时，双尾概率 P 为单尾概率 P 的两倍。

3 df 增大， t 分布接近正态分布，即 t 值接近 u 值。

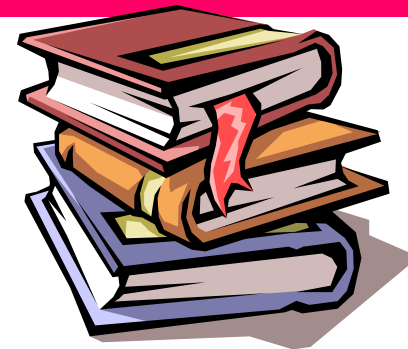
附表 3 t 值表(双尾)

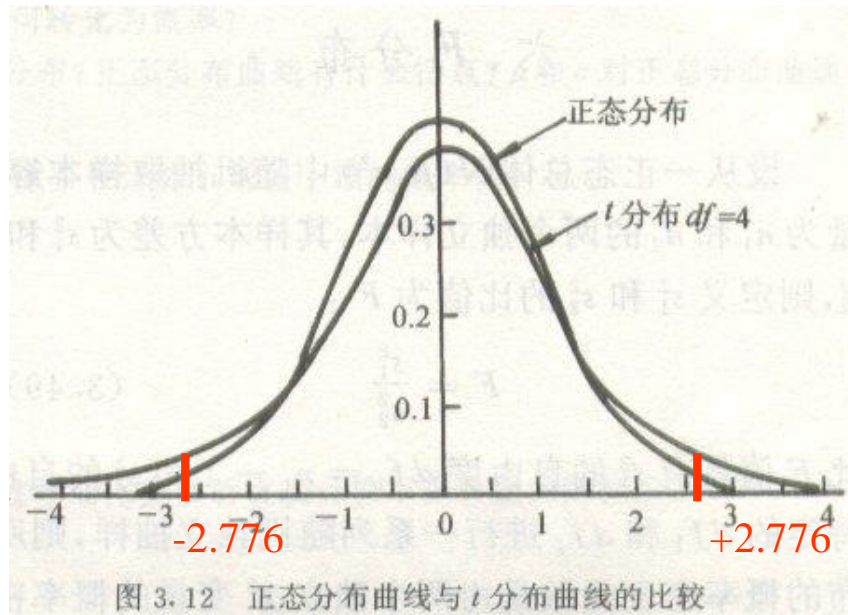
自由度 (df)	概 率 值 (P)								
	0.500	0.400	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	1.000	1.376	3.078	6.314	12.706	25.452	63.657		
2	0.816	1.061	1.886	2.920	4.303	6.205	9.925	14.089	31.598
3	0.765	0.978	1.638	2.353	3.182	4.176	5.841	7.453	12.941
4	0.741	0.941	1.533	2.132	2.776	3.495	4.604	5.598	8.610
5	0.727	0.920	1.476	2.015	2.571	3.163	4.032	4.773	6.859
6	0.718	0.906	1.440	1.943	2.447	2.969	3.707	4.317	5.959
7	0.711	0.896	1.415	1.895	2.365	2.841	3.499	4.029	5.405
8	0.706	0.889	1.397	1.860	2.306	2.752	3.355	3.832	5.041
9	0.703	0.883	1.383	1.833	2.262	2.685	3.250	3.690	4.781
10	0.700	0.879	1.372	1.812	2.228	2.634	3.169	3.581	4.587
11	0.697	0.876	1.363	1.796	2.201	2.593	3.106	3.497	4.437
12	0.695	0.873	1.356	1.782	2.179	2.560	3.056	3.428	4.318
13	0.694	0.870	1.350	1.771	2.160	2.533	3.012	3.372	4.221
14	0.692	0.868	1.345	1.761	2.145	2.510	2.977	3.326	4.140
15	0.691	0.866	1.341	1.753	2.131	2.490	2.947	3.286	4.073
16	0.690	0.865	1.337	1.746	2.120	2.473	2.921	3.252	4.015
17	0.689	0.863	1.333	1.740	2.110	2.458	2.898	3.222	3.965
18	0.688	0.862	1.330	1.734	2.101	2.445	2.878	3.197	3.922
19	0.688	0.861	1.328	1.729	2.093	2.433	2.861	3.174	3.883
20	0.687	0.860	1.325	1.725	2.086	2.423	2.845	3.153	3.850
21	0.686	0.859	1.323	1.721	2.080	2.414	2.831	3.135	3.819
22	0.686	0.858	1.321	1.717	2.074	2.406	2.819	3.119	3.792
23	0.685	0.858	1.319	1.714	2.069	2.398	2.807	3.104	3.767
24	0.685	0.857	1.318	1.711	2.064	2.391	2.797	3.090	3.745
25	0.684	0.856	1.316	1.708	2.060	2.385	2.787	3.078	3.725
26	0.684	0.856	1.315	1.706	2.056	2.379	2.779	3.067	3.707
27	0.684	0.855	1.314	1.703	2.052	2.373	2.771	3.056	3.690
28	0.683	0.855	1.313	1.701	2.048	2.368	2.763	3.047	3.674
29	0.683	0.854	1.311	1.699	2.045	2.364	2.756	3.038	3.659
30	0.683	0.854	1.310	1.697	2.042	2.360	2.750	3.030	3.646
40	0.681	0.851	1.303	1.684	2.021	2.329	2.704	2.971	3.551
60	0.679	0.848	1.296	1.671	2.000	2.299	2.660	2.915	3.460
80	0.678	0.847	1.293	1.665	1.989	2.284	2.638	2.887	3.415
120	0.677	0.845	1.289	1.658	1.980	2.270	2.617	2.860	3.373
∞	0.675	0.842	1.282	1.645	1.960	2.241	2.576	2.807	3.291

$$P(t \geq 2.228) + P(t \leq -2.228) = 0.05$$

$$P(t \leq -1.812) = 0.05$$

$$P(t \geq 1.812) = 0.05$$





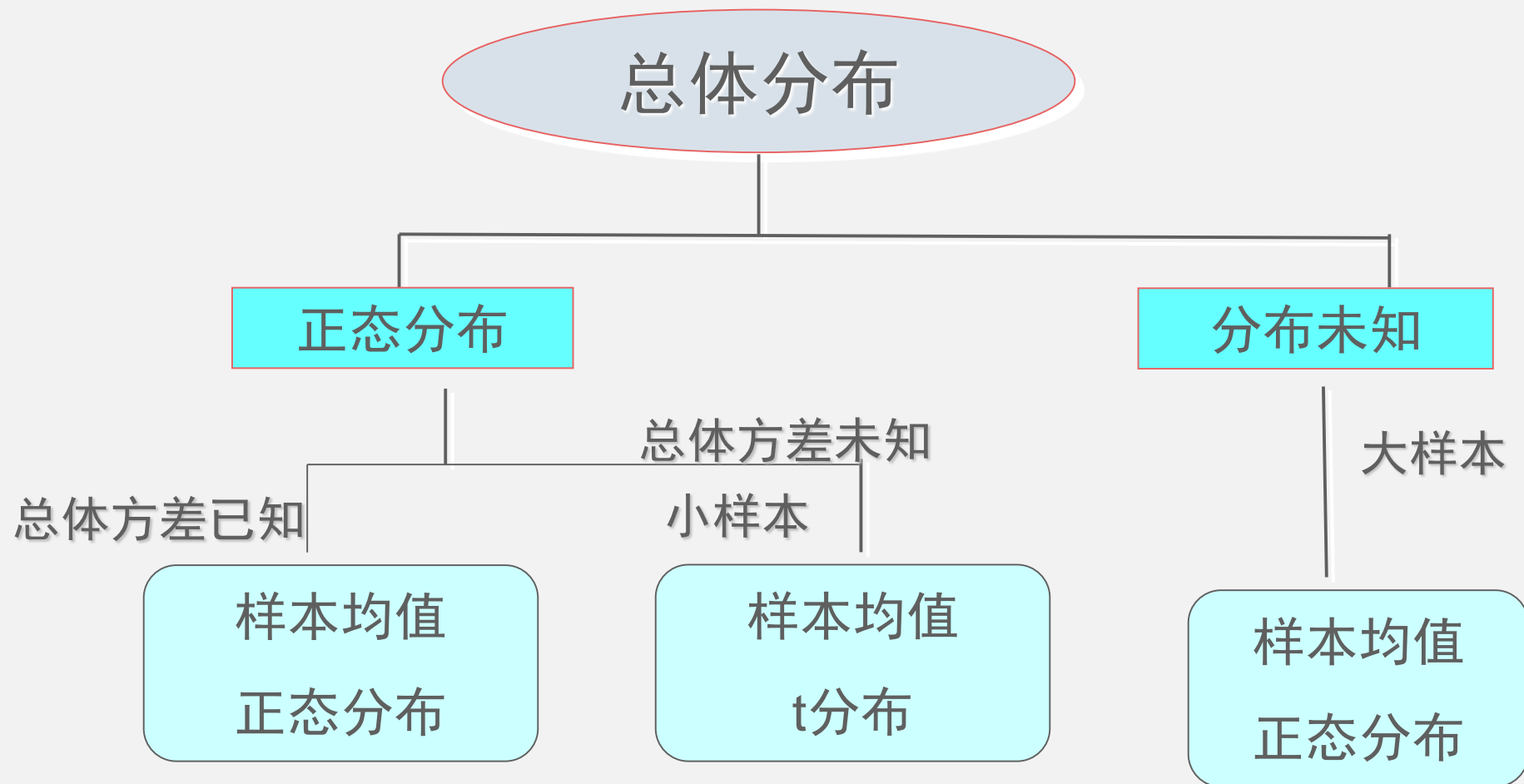
t 落于 $[-t_{0.05}, +t_{0.05}]$ 内的概率为 0.95

t 落于 $[-t_{0.01}, +t_{0.01}]$ 内的概率为 0.99

置信度为 5 % 和 1 % 的 t 临界值。 $t_{0.05 (4)} = 2.776$

$$t_{0.01 (4)} = 4.604$$

抽样分布与总体分布的关系





3.2 样本方差的分布

从方差为 σ^2 的正态总体中，随机抽取 k 个独立样本，计算出样本方差 S^2 ，研究其样本方差的分布。

在研究样本方差的分布时，通常将其标准化，得到 k 个正态离差 u ，则

$$u = \frac{x - \mu}{\sigma} \Rightarrow N(0,1)$$

$$\begin{aligned} \chi^2 &= u_1^2 + u_2^2 + u_3^2 + \dots + u_k^2 = \sum_{i=1}^k u_i^2 \quad \text{df} = k-1 \\ &= \sum_1^k \left(\frac{x - \mu}{\sigma} \right)^2 = \frac{\sum (x - \mu)^2}{\sigma^2} = \frac{dfs^2}{\sigma^2} \end{aligned}$$

样本的分布

从 $N(\mu, \sigma^2)$ 中以 n 为样本容量进行抽样，抽取的样本标准差 s 为连续型随机变量，当我们以

$\frac{(n-1)s^2}{\sigma^2}$ 作为一个新的随机变量时，

称该随机变量为 s^2 的 **标准化** 的随机变量，且该随机变量仍为连续型的随机变量。

我们以 χ^2 来命名新的随机变量，则有：

$$\chi_{df}^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{df \cdot s^2}{\sigma^2}$$

称上式为具有 $n-1$ 自由度的卡方。

χ^2 分布是概率分布曲线随自由度 df 而改变的一类分布:

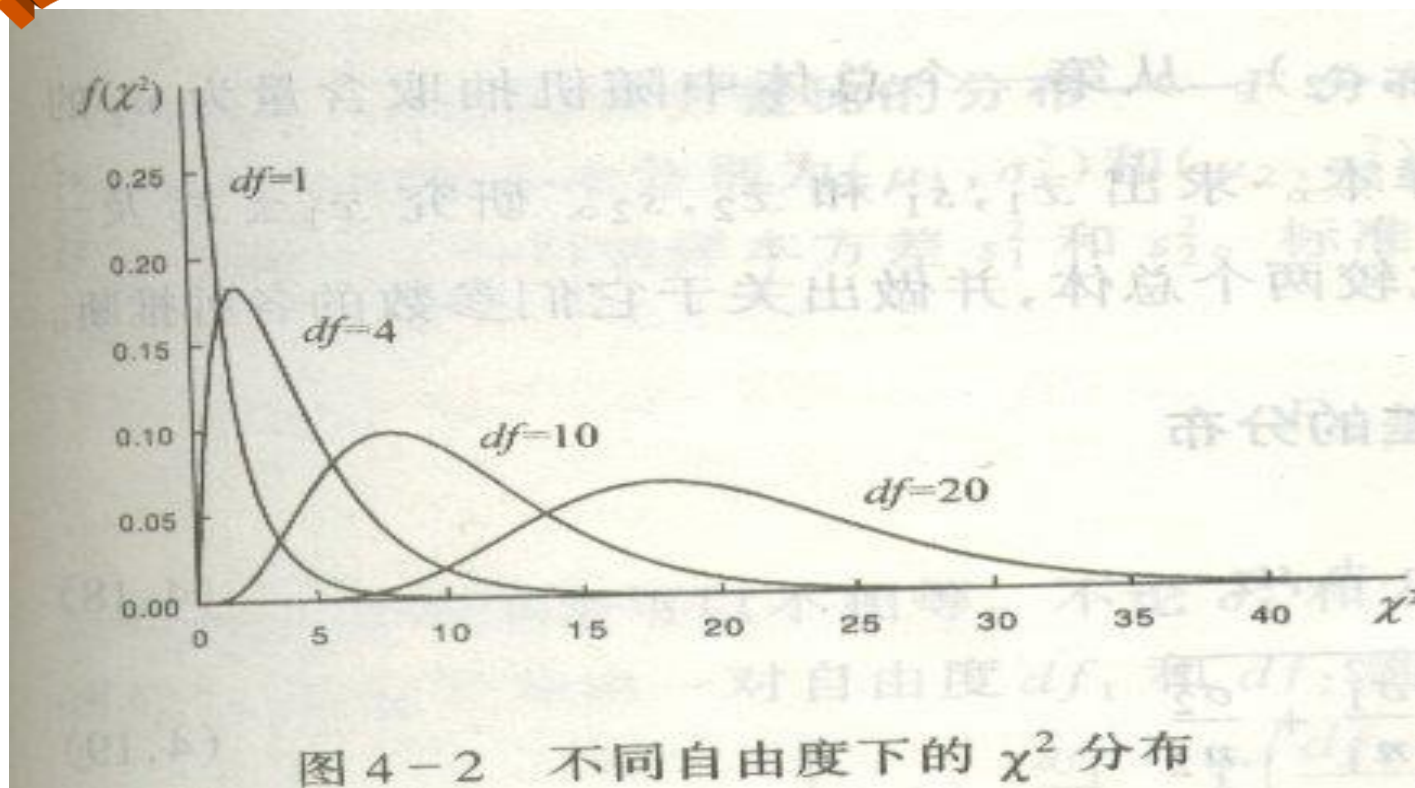
概率密度函数

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

概率累积函数

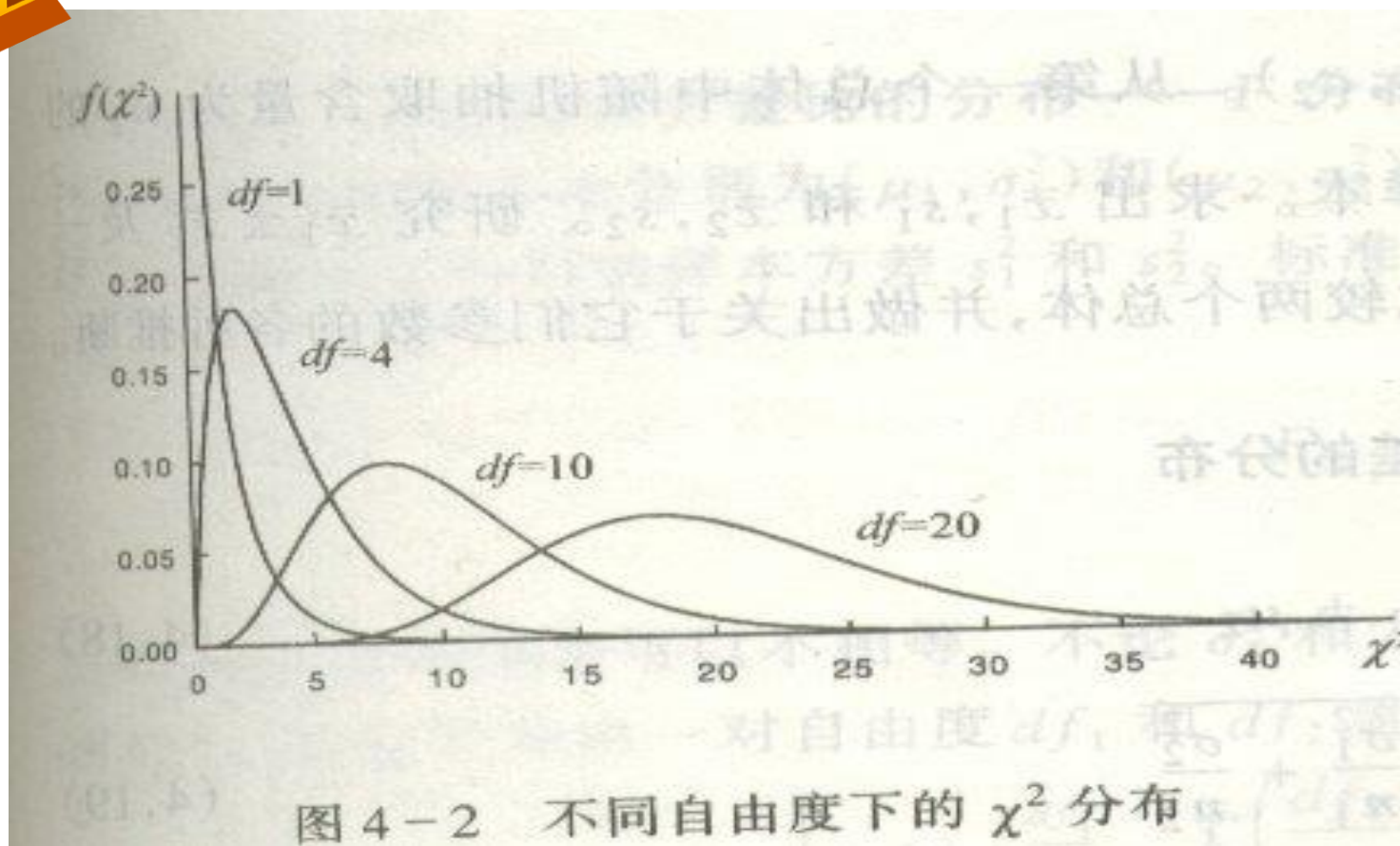
$$F(\chi^2) = \int_0^{\chi^2} f(\chi^2) d(\chi^2)$$

特征



1 χ^2 分布于区间 $[0, +\infty)$ ，并且呈反J型的偏斜分布。

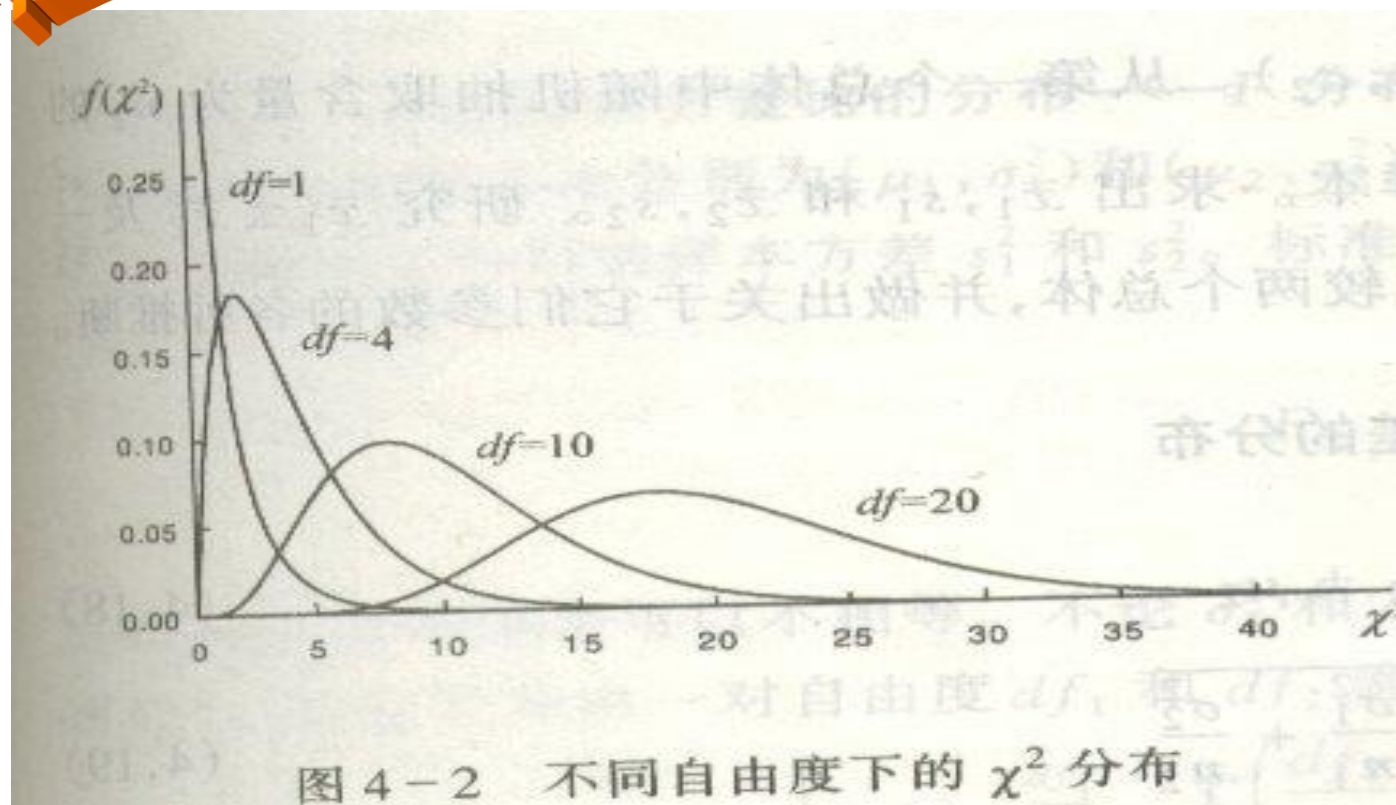
特征



2

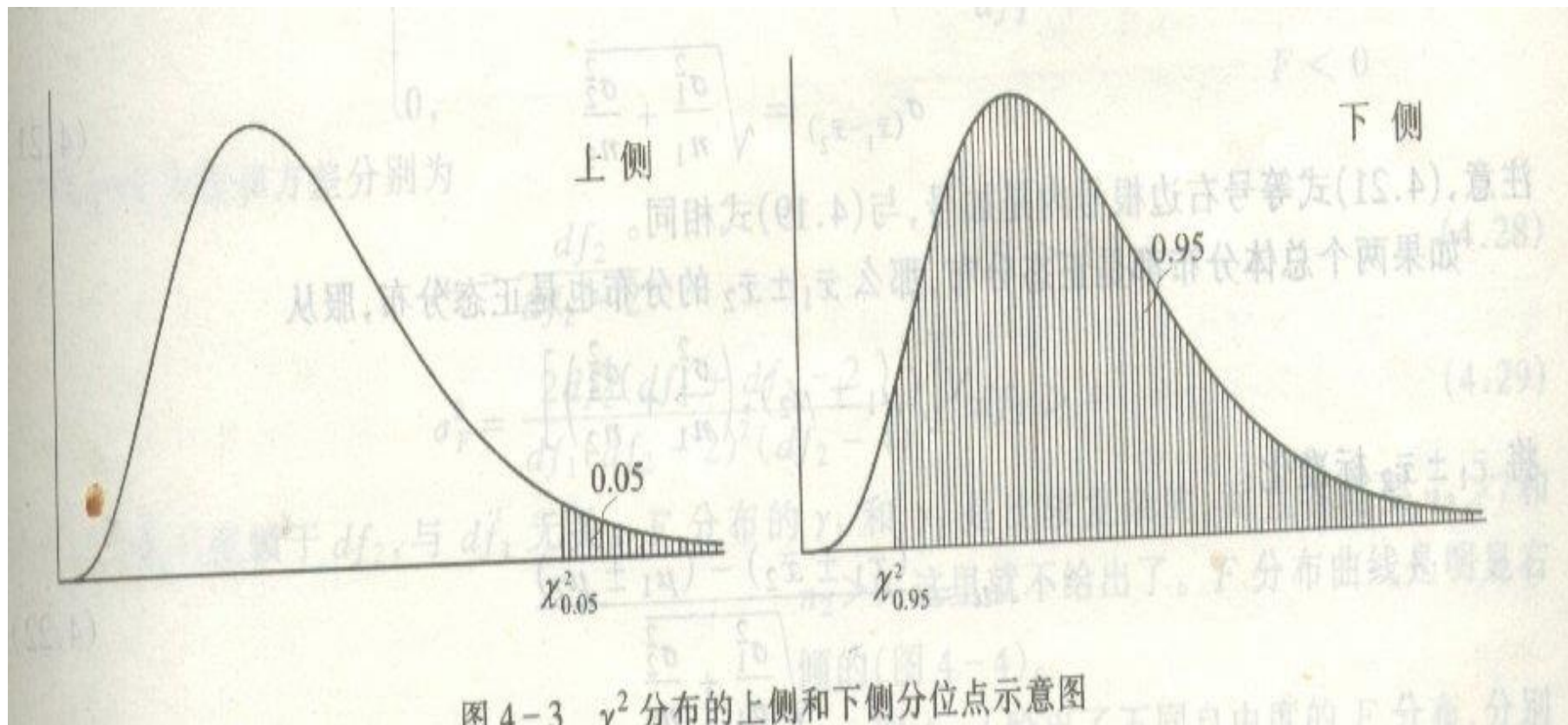
χ^2 分布的偏斜度随自由度降低而增大，当自由度 $df=1$ 时，曲线以纵轴为渐近线。

特征



3

随自由度 df 的增大， χ^2 分布曲线渐趋左右对称，当 $df>30$ 时，卡方分布已接近正态分布。



对于给定的 $\alpha(0<\alpha<1)$,
 称满足条件 $P\{x^2>x_{\alpha}^2(n)\}=\alpha$ 的点 $x_{\alpha}^2(n)$ 为
 x^2 分布的上 α 分位点（右尾概率）。

附表 4 χ^2 值表(右尾)

自 由 度 (<i>df</i>)	概 率 值 (<i>P</i>)												
	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005
1					0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
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	0.30	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	2.20	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	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	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	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	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	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	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	21.60	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	27.20	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	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	49.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67

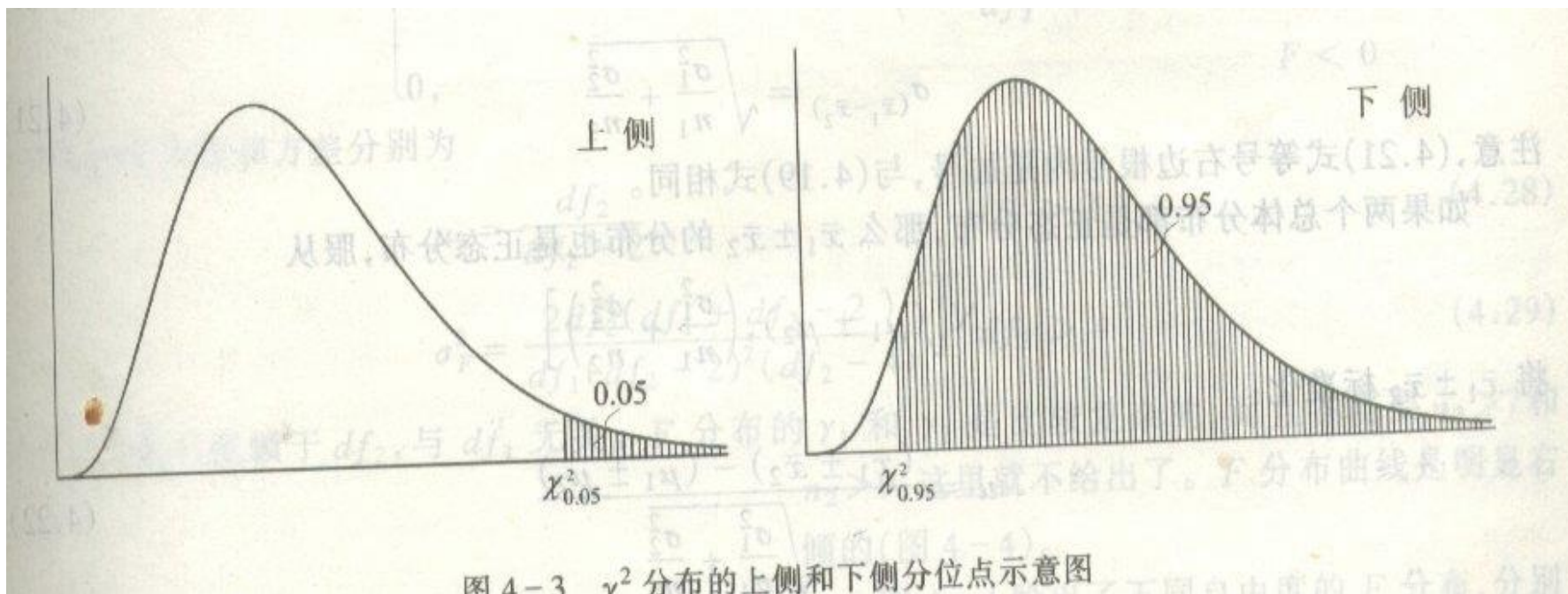
表中表头的概率 α 是 χ^2 大于表内所列 χ^2 值的概率。

$$P(\chi^2 \geq 5.99) = 0.05$$

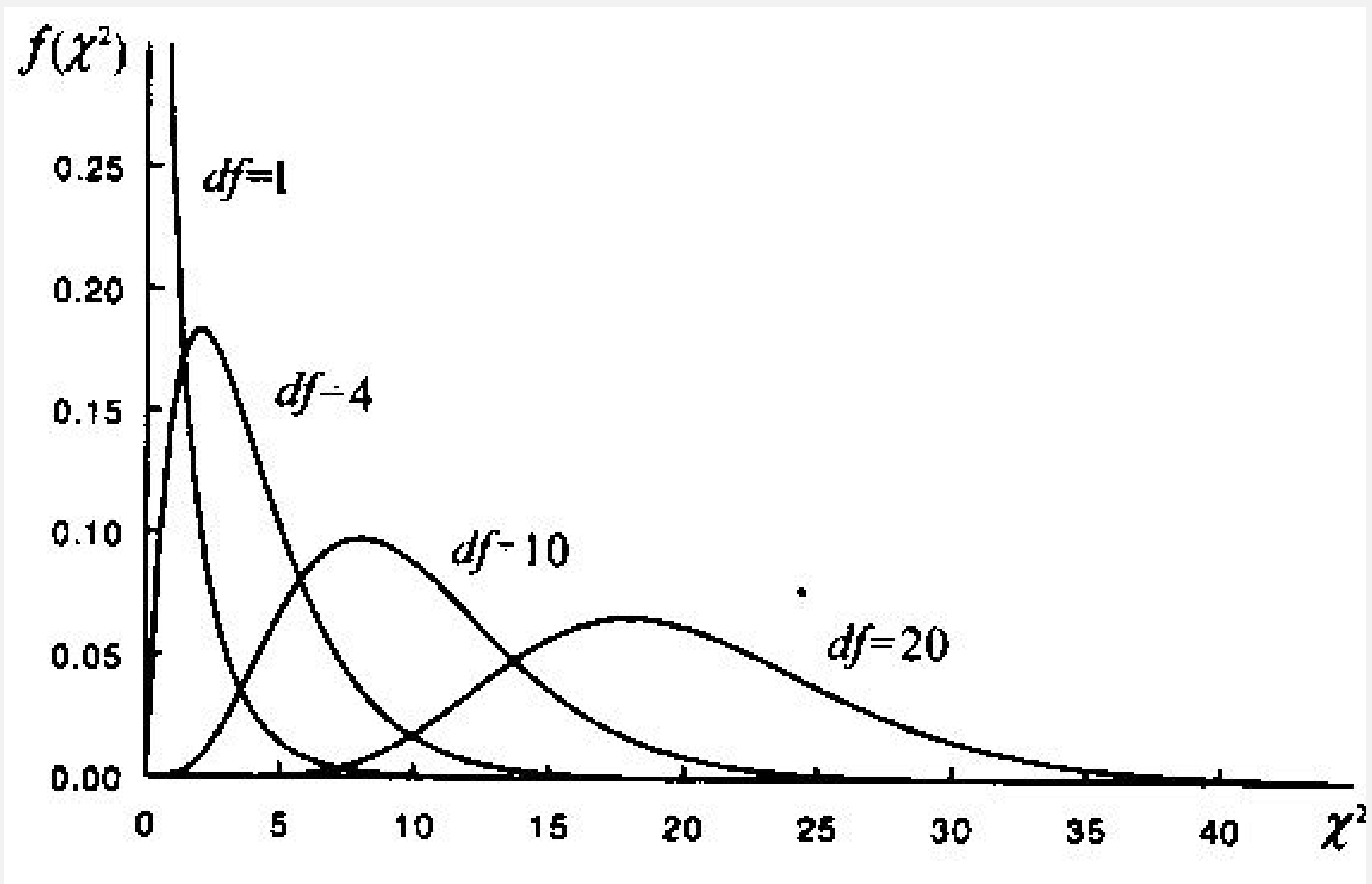
df = 2

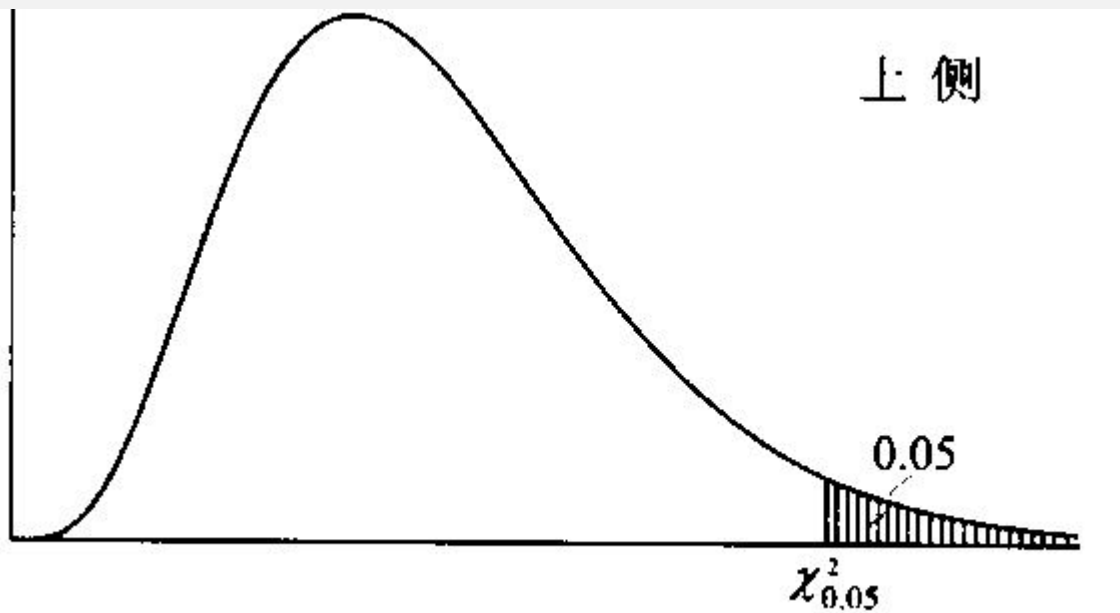
$$P(\chi^2 \geq 9.21) = 0.01$$

$$P(\chi^2 \geq 0.10) = 0.95$$

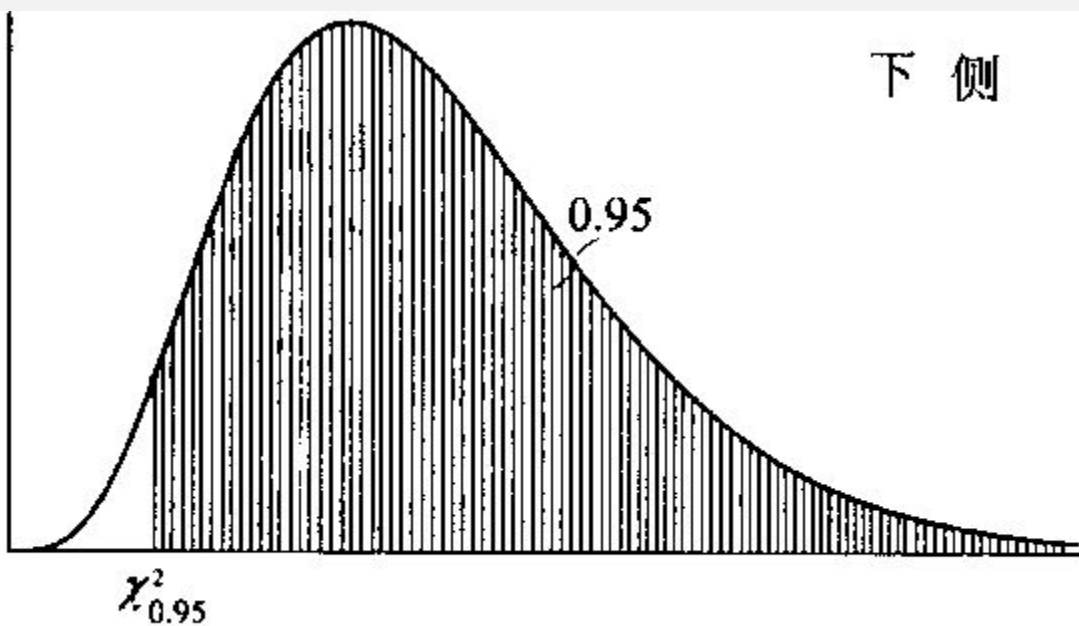


χ^2 分布的分布曲线及分位数





$$P(\chi^2_{df} > \chi^2_{\alpha}) = \alpha$$



$$P(\chi^2_{df} > \chi^2_{1-\alpha}) = 1 - \alpha$$

χ^2 分布

性质

χ^2 分布随机变量的取值范围为 $(0, \infty)$

χ^2 分布 $Y \sim \chi^2(n)$, 则期望 $E(Y)=n$, 均方 $\text{var}(Y) = 2n$

若 $Y_1 \sim \chi^2(n)$, $Y_2 \sim \chi^2(m)$, 且相互独立, 则

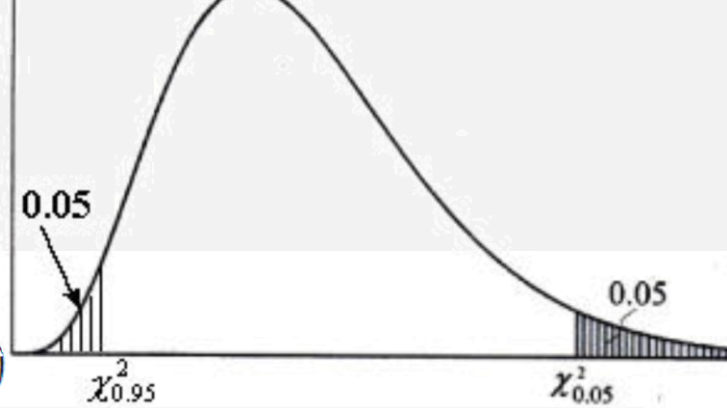
$$Y_1 \pm Y_2 \sim \chi^2(n \pm m)$$

χ^2 分布为非对称分布, 其分布曲线的形状由自由度决定, 自由度越大, 分布越趋于对称

当 $n \rightarrow \infty$, $\chi^2(n) \rightarrow N(n, 2n)$

χ^2 分布上侧分位数表

$$P(X \geq \chi^2_{\alpha}) = \alpha$$



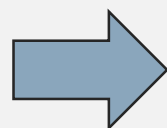
附表三 χ^2 分布上侧分位数表

$$(P\{\chi^2(n) > \chi^2_{\alpha}(n)\} = \alpha)$$

α n	0.995	0.99	0.975	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.025	0.01	0.005
1	0.00004	0.00016	0.001	0.004	0.016	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	7.584	10.341	13.701	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	8.438	11.340	14.845	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	9.299	12.340	15.984	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	10.165	13.339	17.117	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	11.037	14.339	18.245	22.307	24.996	27.488	30.578	32.801

知识小结

总体 $X \sim N(\mu, \sigma^2)$



样本平均数

总体方差已知

$$u = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

总体方差未知
(小样本)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

样本方差

$$\chi^2_{df} = \frac{(n-1)s^2}{\sigma^2} = \frac{df \cdot s^2}{\sigma^2}$$



3.2 从两个正态总体中抽取的样本统计量的分布

1. 标准差 σ_i 已知，两个平均数的和与差的分布

两个正态总体分别为： $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ ，分别进行独立随机抽样含量为 n_1 和 n_2 的样本，则两个样本样本平均数的和与差的分布为：

$\bar{y}_1 \pm \bar{y}_2$ 服从均值为 $(\mu_1 \pm \mu_2)$ ，方差为 $(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)$ 的正态分布，记作：

$$\bar{y}_1 \pm \bar{y}_2 \text{服从 } N(\mu_1 \pm \mu_2, \sigma_1^2 / n_1 + \sigma_2^2 / n_2)$$

将 $\bar{y}_1 \pm \bar{y}_2$ 标准化, 则

$$u = \frac{(\bar{y}_1 \pm \bar{y}_2) - (\mu_1 \pm \mu_2)}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

u 服从 $N(0, 1)$

2. 标准差 σ_i 未知但相等，两个平均数的和与差的分布

两个正态总体分别为： $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ ，分别进行独立随机抽样含量为 n_1 和 n_2 的样本。其中 σ_1 与 σ_2 未知，但是 $\sigma_1 = \sigma_2 = \sigma$ 则两个样本样本平均数的和与差的分布为：

$\bar{y}_1 \pm \bar{y}_2$ 服从 $df_1 + df_2$ 自由度的t分布。 $df_1 = n_1 - 1$ ； $df_2 = n_2 - 1$ 。

$$t_{df_1+df_2} = \frac{(\bar{y}_1 \pm \bar{y}_2) - (\mu_1 \pm \mu_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

如果 $n_1=n_2$,

$$t_{2n-2} = \frac{(\bar{y}_1 \pm \bar{y}_2) - (\mu_1 \pm \mu_2)}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$$

3. 两个样本方差比的分布 – F分布

两个正态总体分别为： $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ ，分别进行独立随机抽样含量为 n_1 和 n_2 的样本。标准化的样本方差比的分布称为F分布

$$F_{df_1, df_2} = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

F 分布的概率密度函数是两个独立 χ^2 变量的概率密度所构成的联合概率密度。

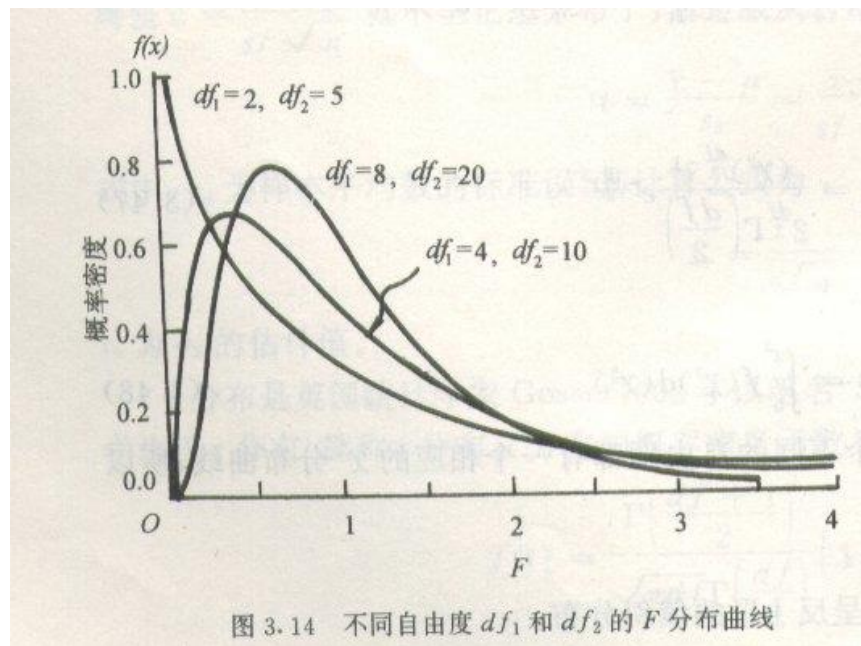
$$f(F) = \frac{\Gamma(\frac{df_1 + df_2}{2})}{\Gamma(\frac{df_1}{2})\Gamma(\frac{df_2}{2})} df_1^{\frac{df_1}{2}-1} df_2^{\frac{df_2}{2}-1} \frac{F^{\frac{df_1}{2}-1}}{(df_1 F + df_2)^{\frac{df_1 + df_2}{2}}}$$

F 分布是随自由度 df_1 和 df_2 进行变化的一组曲线。

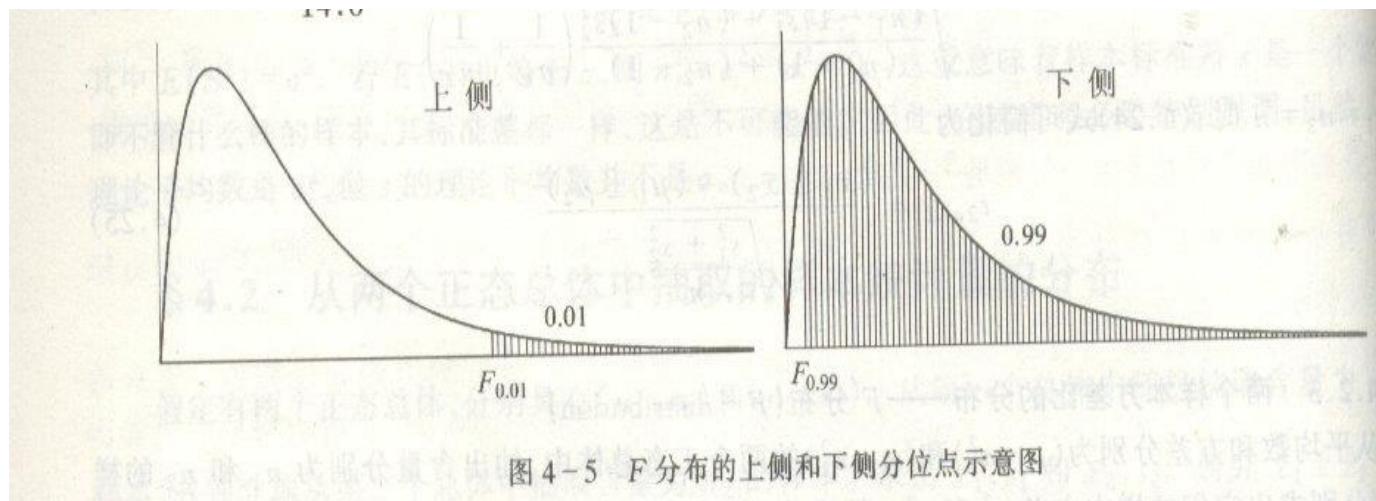
F 分布的概率累积函数

$$F(F) = \int_0^F f(F) dF$$

特征



- 1 F 分布的平均数 $\mu_F=1$ ， F 的取值区间为 $[0, +\infty)$
- 2 F 分布曲线的形状仅决定于 df_1 和 df_2 。在 $df_1=1$ 或 2 时， F 分布曲线呈严重倾斜的反向 J 型，当 $df_1 \geq 3$ 时，转为左偏曲线。



对于给定的 $\alpha(0 < \alpha < 1)$ 称满足条件

$P\{F > F_{\alpha}(n_1, n_2)\} = \alpha$ 的点 $F_{\alpha}(n_1, n_2)$ 为
F 分布的上 α 分位点（或临界值点）。

附表 5 F 值表(右尾) $P = 0.05$

df_2	df_1 (大 方 差 自 由 度)															
	1	2	3	<u>4</u>	5	6	7	8	9	10	12	14	16	18	20	
1	161	200	216	225	230	234	237	239	241	242	244	245	246	247	248	
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.41	19.42	19.43	19.44	19.44	
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	5.80	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56	
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87	
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44	
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15	
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94	
<u>10</u>	4.96	4.10	3.71	<u>3.48</u>	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77	

$$P(F \geq 3.48)$$

$$= 0.05$$

$$F_{0.05(4,10)} = 3.48$$

 $P = 0.01$

df_2	df_1 (大 方 差 自 由 度)															
	1	2	3	<u>4</u>	5	6	7	8	9	10	12	14	16	18	20	
1	405	500	540	563	576	586	593	598	602	606	611	614	617	619	621	
2	98.49	99.00	99.17	99.25	99.30	99.33	99.34	99.36	99.38	99.40	99.42	99.43	99.44	99.44	99.45	
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.05	26.92	26.83	26.75	26.69	
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.37	14.24	14.15	14.07	14.02	
5	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.89	9.77	9.68	9.61	9.55	
6	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.93	7.87	7.72	7.60	7.52	7.45	7.40	
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.36	6.27	6.21	6.16	
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.56	5.48	5.41	5.36	
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	5.00	4.92	4.86	4.81	
<u>10</u>	10.04	7.56	6.55	<u>5.99</u>	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.60	4.52	4.46	4.41	

$$P(F \geq 5.99)$$

$$= 0.01$$

$$F_{0.01(4,10)} = 5.99$$

- 我们所研究的抽样分布，全部都是建立在正态分布基础之上的。或者说，全部都是从正态分布的总体中进行抽样的。但是根据中心极限定理，从一个非正态的总体中抽取的容量为 n 的样本，当充分大时，样本平均数渐近服从正态分布。因此平均数抽样对总体正态性的要求并不十分严格，但方差的抽样分布对正态性的要求是十分严格的。

统计量的极限分布

- 当样本大小趋于无穷时，若统计量的分布趋于一定的分布，测称这个分布为统计量的**极限分布或渐近分布**，也称**大样本分布**. 这可理解为：当样本大小很大时，统计量的**近似分布**.
- 意义：(1)弄清一个统计推断方法的优良性如何，甚至单纯为了一实现统计推断，往往有必要知道统计量的分布，但统计量的分布一般很难求出，建立其极限分布就提供了一种近似解法的可能性. (2)统计推断方法的某些优良性准则，本身就是建立在样本大小趋于无穷的基础上.
- 当样本大小趋于无穷时，一个统计量或统计推断方法的性质，称为**大样本性质**. 大样本性质只在在样本大小趋于无穷时才有意义. 与此相对，一个统计量或统计推断方法的性质，**如果在样本大小固定时有意义，就称为是小样本性质**.
- 大样本和小样本的差别不在于样本个数的多少，而在于样本是在样本大小 $n \rightarrow \infty$ 去讨论，还是 n 固定时讨论.



大数定律和中心极限定理

大数定律

大数定律：是概率论中用来阐述大量随机现象**平均结果**稳定性的一系列定律的总称。随着样本容量 n 的增加，总体 X 的随机样本的平均数 $EX = \mu$ 的偏差比任何指定的数都小的概率趋向于1.

样本容量越大，样本统计量与总体参数之差越小。



大数定律

(1) 贝努里大数定律

设 m 是 n 次独立试验中事件 A 出现的次数，而 p 是事件 A 在每次试验中出现的概率，则对于任意小的正数 ε ，有如下关系：

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{m}{n} - p \right| < \varepsilon \right\} = 1$$



大数定律

(2) 辛钦大数定律

设 $x_1, x_2, x_3, \dots, x_n$ 是来自同一总体的变量,
 $EX = \mu$, 则对于给定的 $\varepsilon > 0$, 将有

$$\lim_{n \rightarrow \infty} P \{ |\bar{x} - \mu| < \varepsilon \} = 1$$



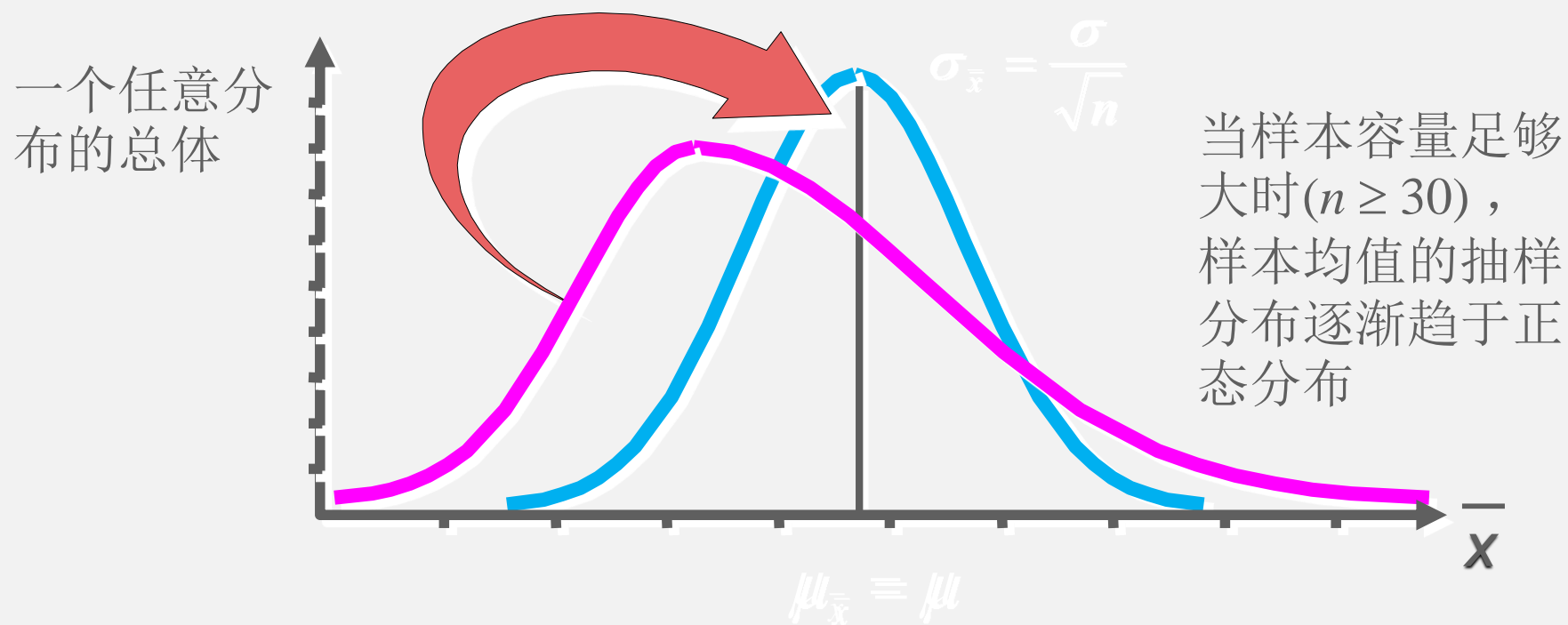
中心极限定理(central limit theorem)

两个定理：

- 1.若随机变量 x 服从正态分布 $N(\mu, \sigma^2)$ ； x_1 、 x_2 、 \cdots 、 x_n 是由 x 总体得来的随机样本,则统计量 \bar{X} 的概率分布也是正态分布,且服从正态分布 $N(\mu, \sigma^2 / n)$ 。
2. 若随机变量 x 服从均值为 μ ,方差是 σ^2 的($EX = \mu$, $DX = \sigma^2$ 存在, 且 $\sigma^2 \neq 0$) 分布； x_1 、 x_2 、 \cdots 、 x_n 是 X 的容量为 n 的随机样本, 则当 n 相当大时, 统计量 \bar{X} 的概率分布逼近正态分布 $N(\mu, \sigma^2 / n)$ 。这就是中心极限定理

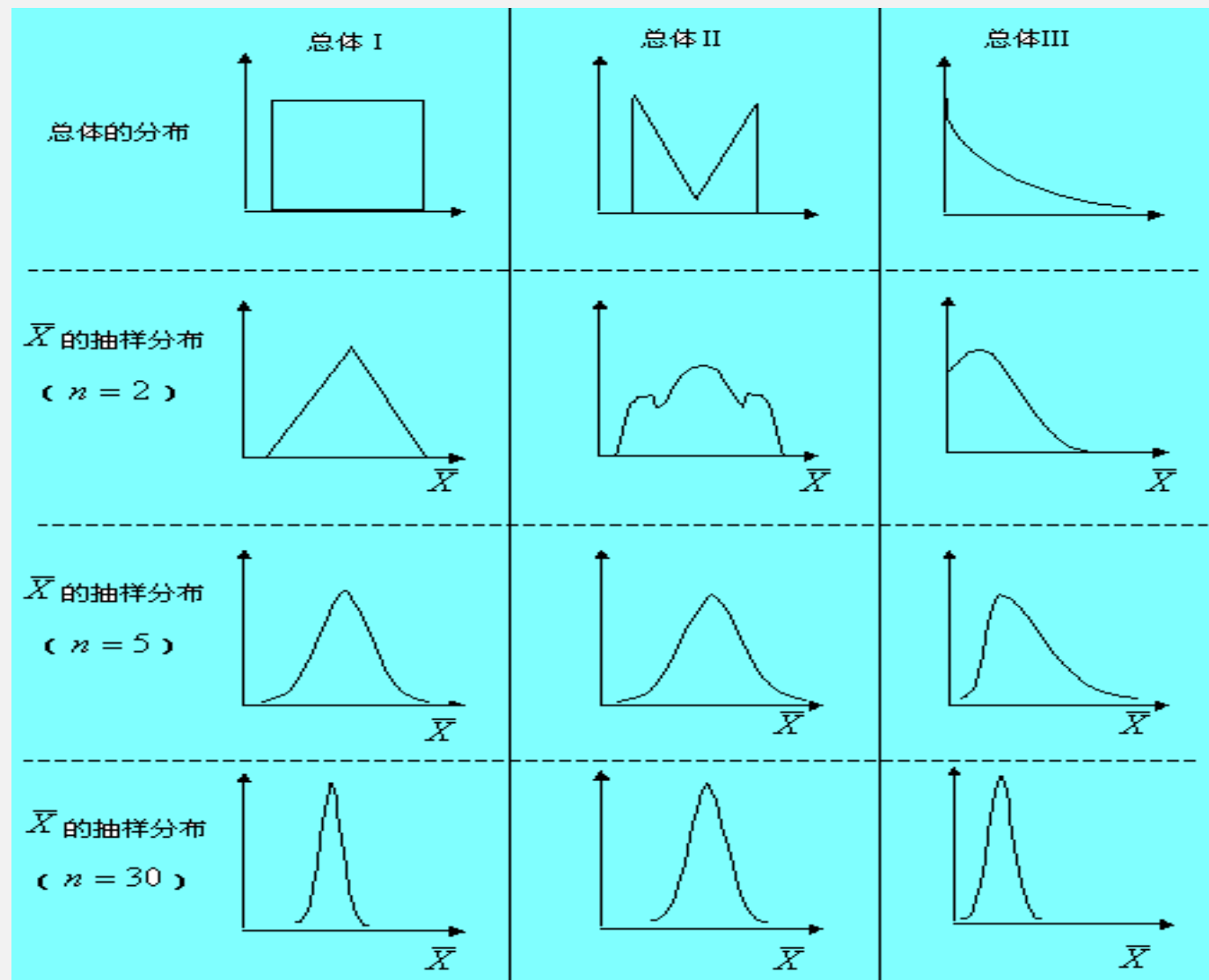
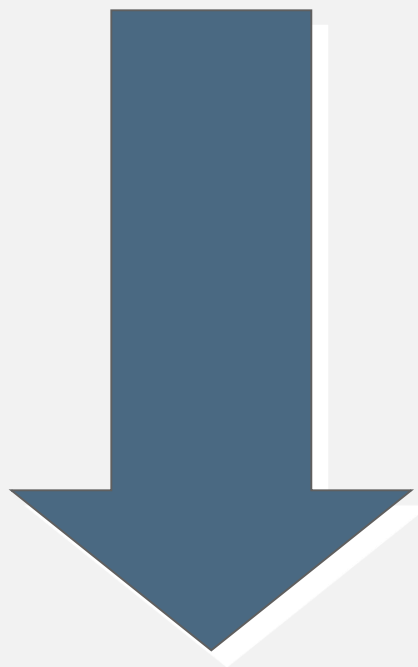
中心极限定理 (central limit theorem)

从均值为 μ ，方差为 σ^2 的一个任意总体中抽取容量为 n 的样本，当 n 充分大时，样本均值的抽样分布近似服从均值为 μ 、方差为 σ^2/n 的正态分布



中心极限定理 (central limit theorem)

\bar{X} 的分布趋于正态分布的过程



中心极限定理告诉我们：无论一个总体的分布如何，只要它有有限的方差，那么当 n 相当大时，容量为 n 的一个随机样本的平均数将近似的服从正态分布。随着样本含量 n 的增大, 样本平均数的分布愈来愈从不连续趋向于连续的正态分布。当 $n > 30$ 时， \bar{X} 的分布就近似正态分布了。

不论 x 变量是连续型还是离散型，也无论 x 服从何种分布，一般只要 $n > 30$ ，就可认为 \bar{x} 的分布是正态的。若 x 的分布不很偏倚，在 $n > 20$ 时， \bar{X} 的分布就近似于正态分布了。