第四章　抽样分布

张爱莲

新疆大学生命科学与技术学院　新疆生物资源基因工程重点实验室

zalxju@gmail. com

20100909

课前提问

1. 正态分布 密度函数 分布曲线

   累积分布函数

2. 标准正态分布 密度函数 分布曲线

3. 正态分布的临界值（分位数）

本章内容

4.1 样本平均数的抽样分布

4.2 从一个正态总体中抽样

4.3 从两个正态总体中抽样

课前提问

- 什么是抽样分布？
- 样本平均数的抽样分布特点？
- 从一个正态总体中抽样，总体σ已知时，样本平均数的分布？

标准正态分布统计量的计算公式？

4.1 样本平均数的抽样分布

- 4.1.1 抽样分布概述
- 4.1.2 样本平均数的抽样分布

4.1.1抽样分布概述

研究总体与从中抽取的样本之间的关系是统计学的中心内容。对这种关系的研究可从两方面着手:

- 一是从总体到样本 –抽样分布的问题;
- 二是从样本到总体–统计推断问题。

三种不同性质的分布

- 总体分布（population distribution)
  就是指总体中各元素的观察值所形成的分布。

- 样本分布(sample distribution)
  一个样本中各观察值的分布。

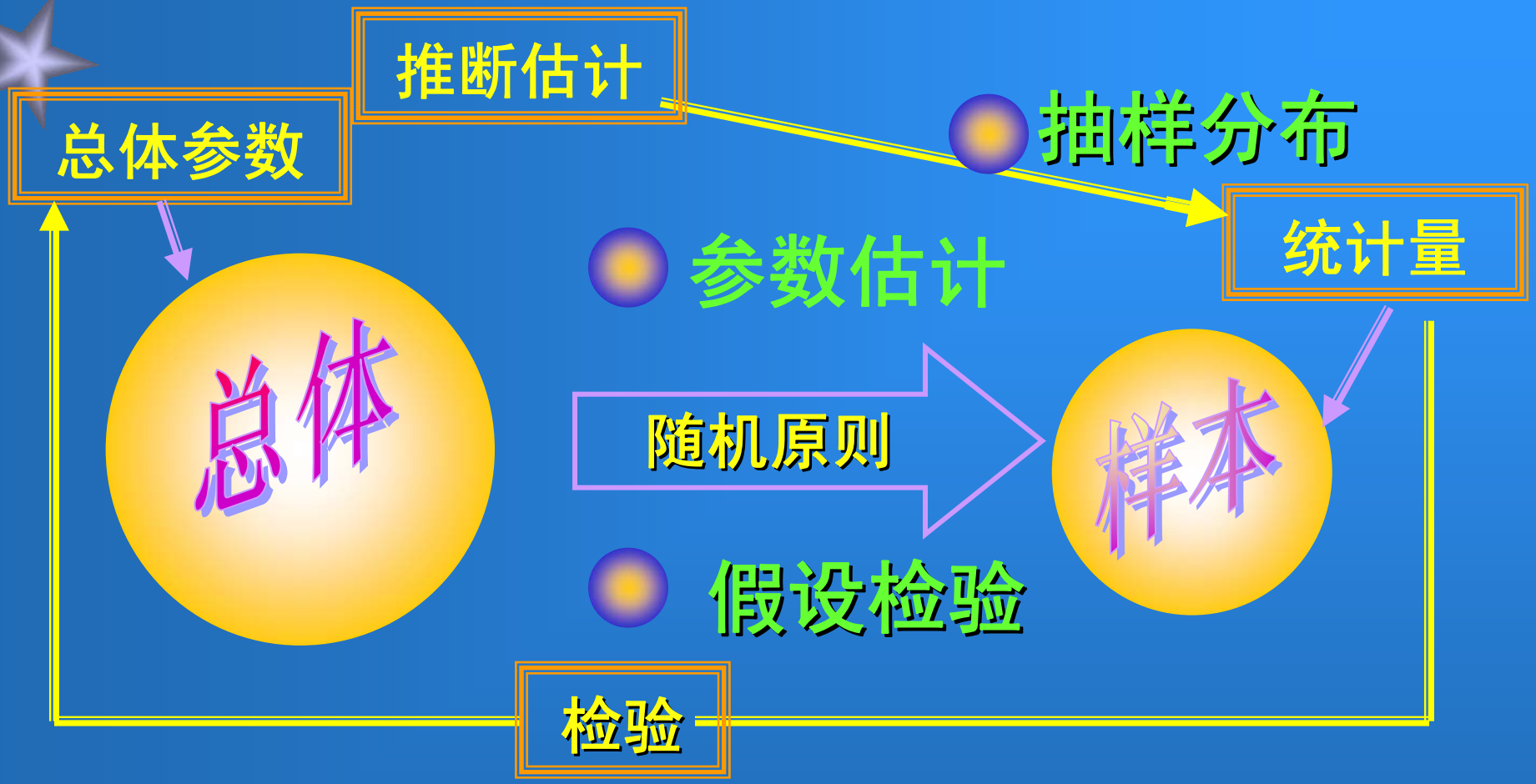
- 抽样分布（sampling distribution)
  样本统计量的概率分布，比如，样本平均值的分布，样本方差的分布等都成为抽样分布。

- 总体参数的值是一个常数，这个常数不会随样本的不同而变化。
- 样本统计量的值高度依赖于所抽取的样本.

不同的样本就会有不同的样本均值。正是因为样本统计量是依据样本而变化的，所以根据统计量来推断总体的参数必然具有某种不确定性。

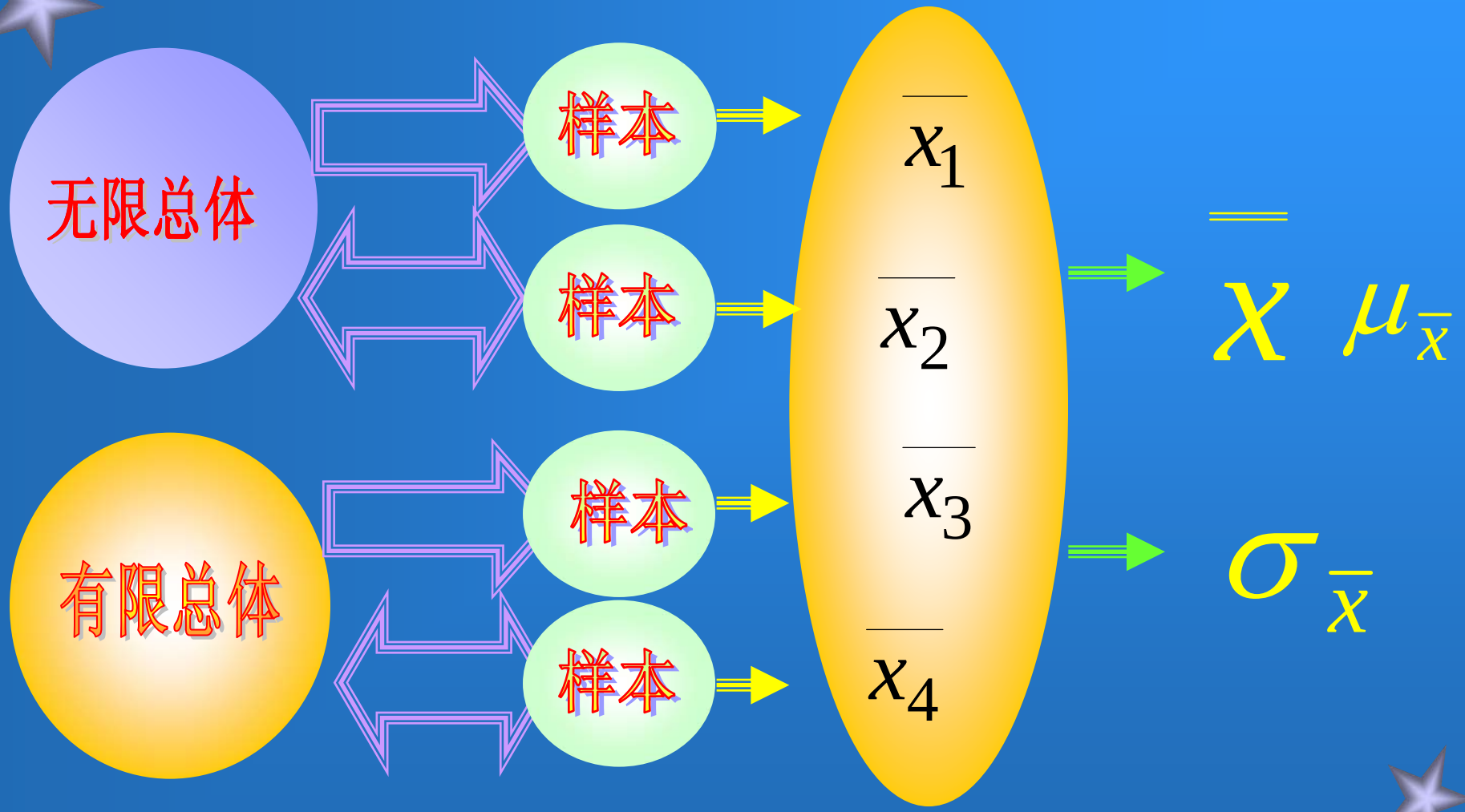
Question: 又如何判断用样本统计量来推断总体的可靠性呢？

- 样本统计量的分布具有某种确定的性质，这些性质对我们来说是已知的。而且反应在它的抽样分布之中。
- 样本统计量的抽样分布构成了我们推断总体参数的理论基础。



★ 统计推断：利用样本统计量对总体某些性质或数量特征进行推断。

抽样分布


无限总体
有限总体
样本
样本
样本
样本
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
$\bar{x}_4$
$\bar{\bar{X}}$ $\mu_{\bar{x}}$
$\sigma_{\bar{x}}$

- 一般意义上说，抽样分布是指样本统计量的概率分布。
- 说的具体一些，某个样本统计量的抽样分布，从理论上说就是在重复选取容量为n的样本时，由每一个样本算出的该统计量数值的相对频数分布。
- 由于现实中我们不可能将所有的样本都抽出来，因此，统计量的抽样分布实际上是一种理论分布。

统计量与抽样分布

•统计量：即样本特征数。

如：
- 样本平均值 $\overline{X} = \frac{\Sigma X_i}{n}$
- 样本方差 $S^2 = \frac{1}{n-1}\Sigma(X_i - \overline{X})^2$

•抽样分布：某一统计量所有可能的样本的取值形成的分布。

性　质
- $0 \leqslant P(X_i) \leq 1$
- $\Sigma P(X_i) = 1$

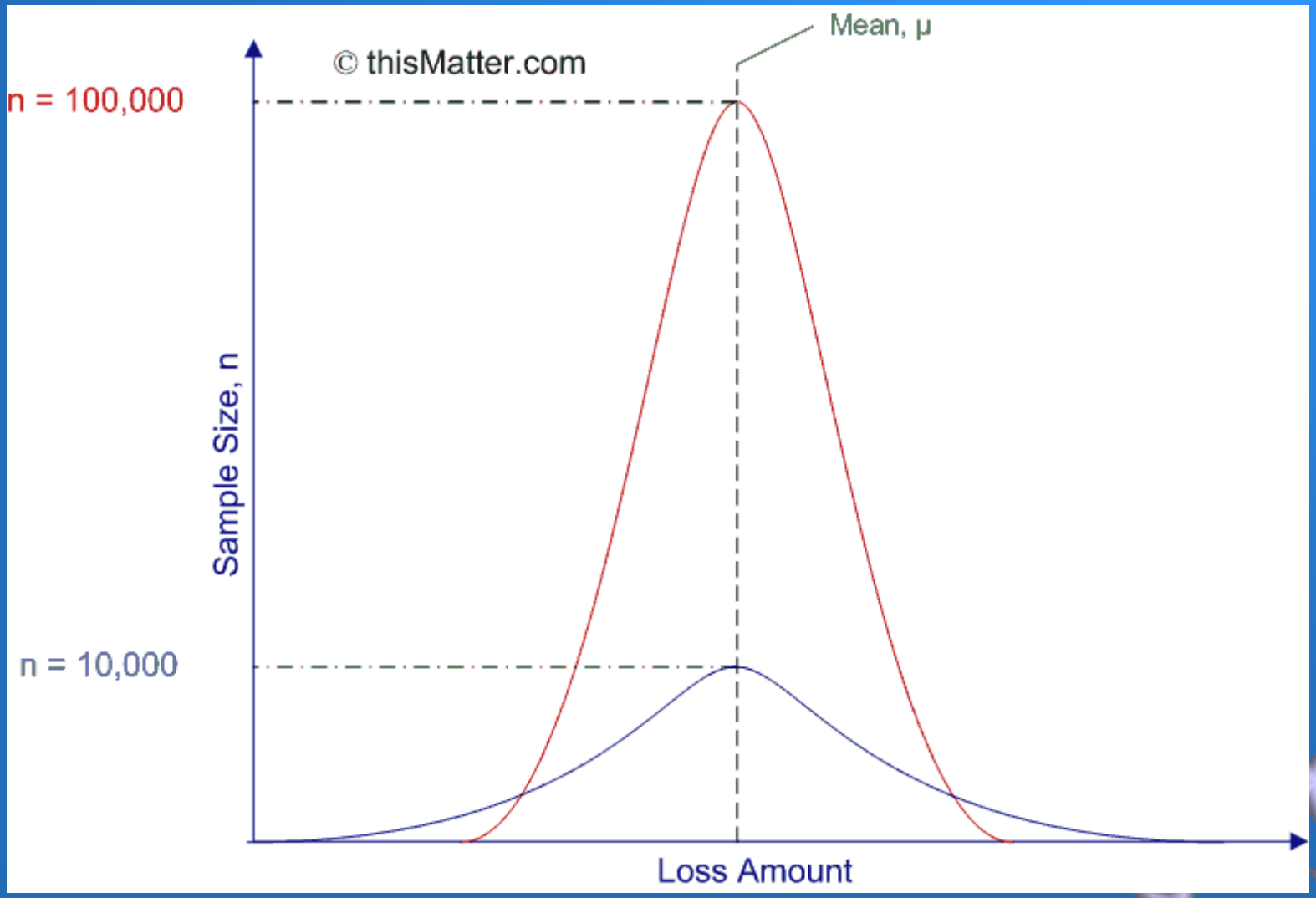
数字特征
- 平均值E（X）
- 方差$E[x-E(x)]^2$

★ 抽样分布的标准差就是推断的抽样误差。

4.1.2 样本平均数的抽样分布

- 设有一个任意总体（μ，σ^2），将此总体称为原总体。现从这个总体中随机抽取含量为n的样本，样本平均数记为 $\overline{X}$ 。
- 从原总体中可抽出N^n个含量为n的样本（有限总体放回式抽样）。这些样本的平均数不尽相同，与原总体平均数μ相比表现出不同程度的差异。这种差异是由随机抽样造成的，称为抽样误差。

样本与总体是什么关系？


Mean, μ
© thisMatter.com
n = 100,000
n = 10,000
Sample Size, n
Loss Amount

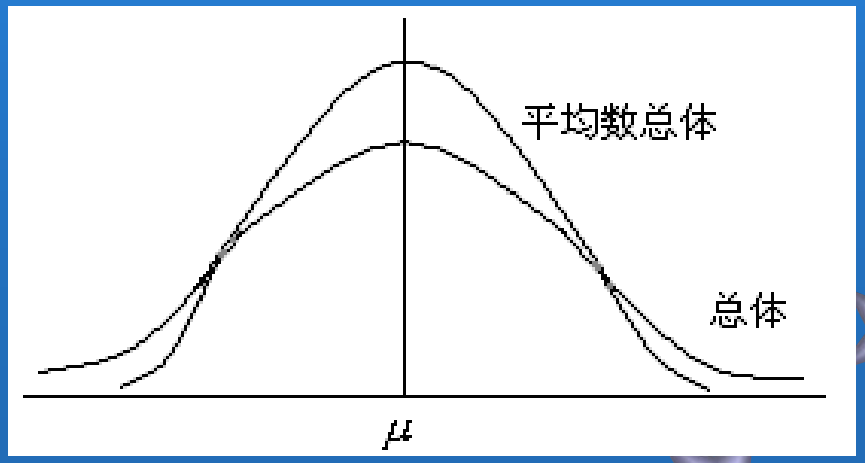
- 样本平均数的概率分布叫做**样本平均数的抽样分布。**
- 由样本平均数构成的总体称为**样本平均数的抽样总体。**其平均数和标准差分别记为 $\mu_{\bar{x}}$和 $\sigma_{\bar{x}}$。

$\sigma_{\bar{x}}$**是样本平均数抽样总体的标准差，简称标准误(standard error)，它表示平均数抽样误差的大小。**

- 抽样总体的两个参数与原总体的两个参数有如下关系：

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

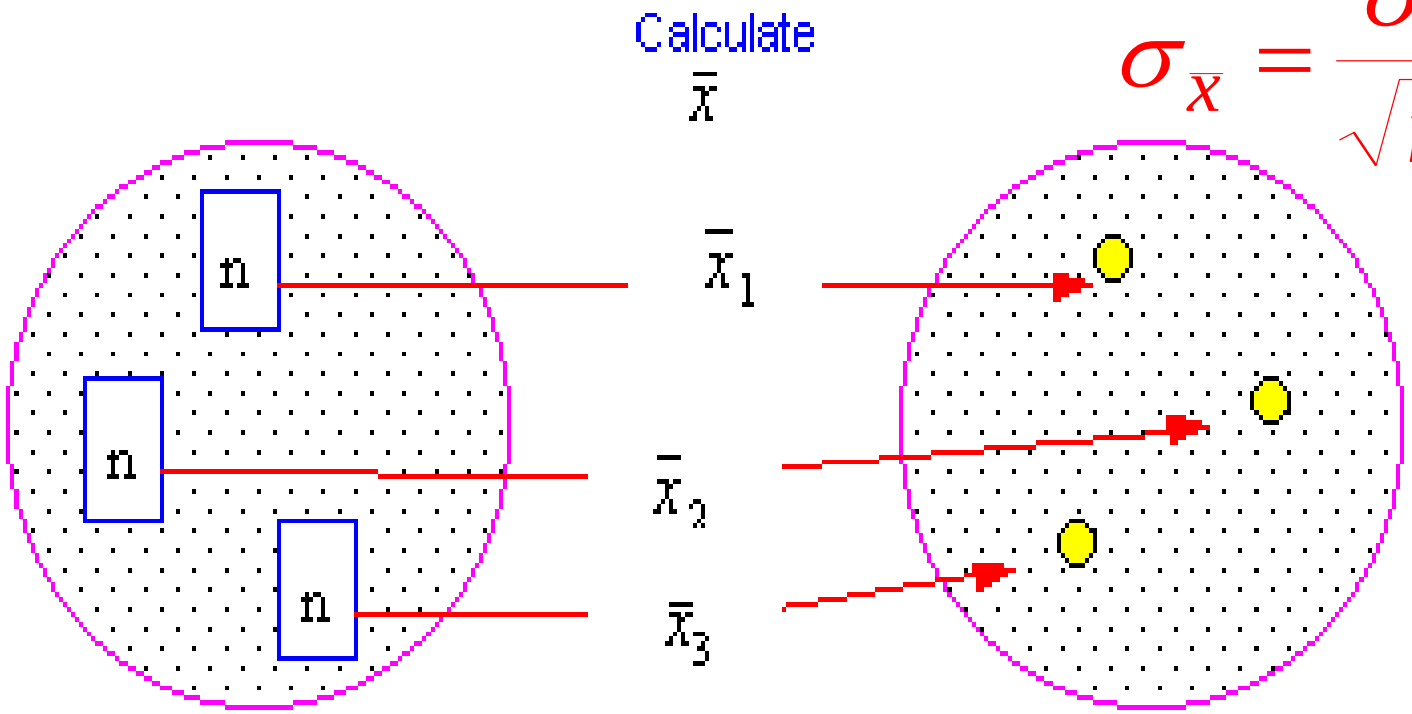
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



Select sample of size = 5 number of children ever born

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

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



Population consisting of 4,171 number of children ever born

Repeat this process an infinite number of times

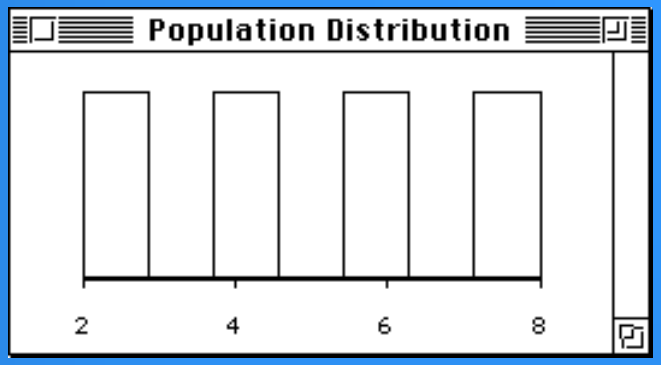
Sampling distribution of $\bar{x}$

样本平均数抽样分布的形成过程

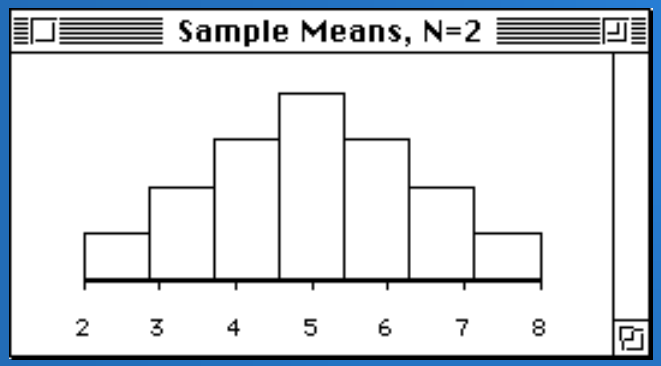
例一：

Population Distribution for [2, 4, 6, 8]

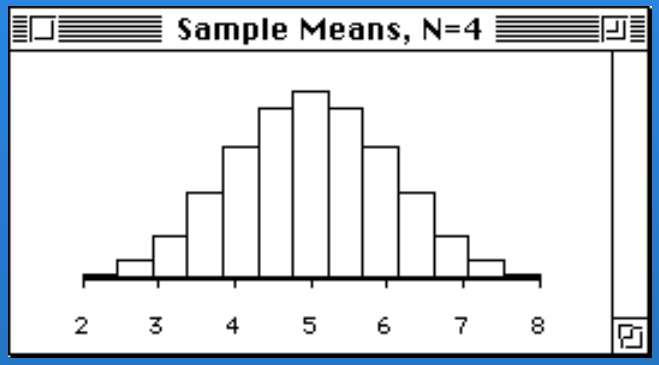
This Population is very non-normal (uniform)



We extend their example by constructing distributions of sample means for sample sizes of N=2, N=4 and N=5.

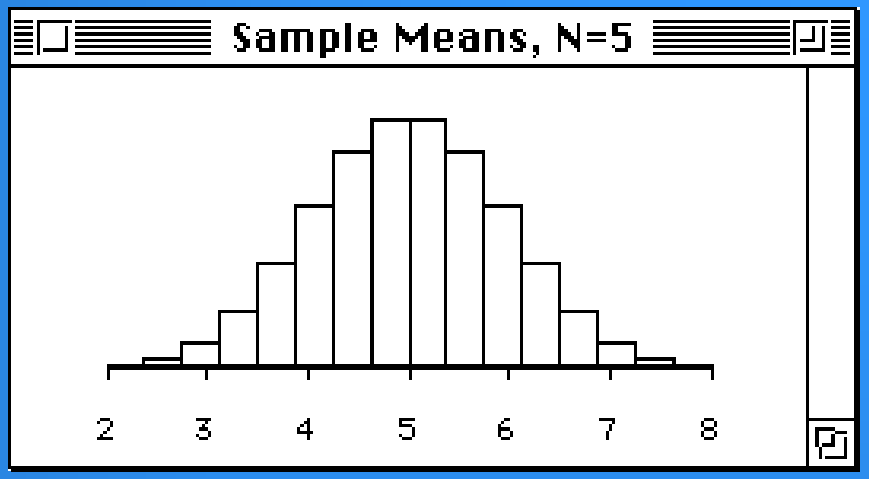


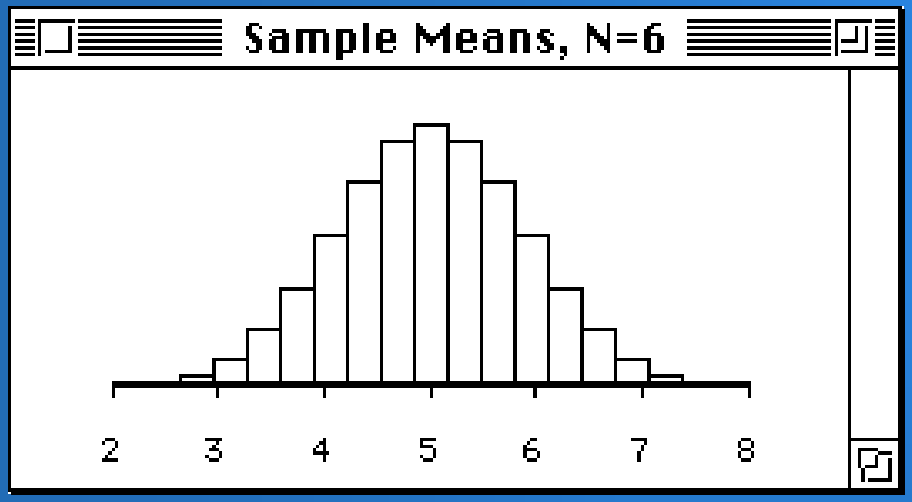
Sampling Distribution of Sample Means (N=2). Normality is better than those above.



Sampling Distribution of Sample Means (N=4). Normality is even better than those above.

Sampling Distribution of Sample Means (N=5) .This distribution is very normal.





Sampling Distribution of Sample Means (N=6) This distribution is even more normal.

To repeat the point of this example:

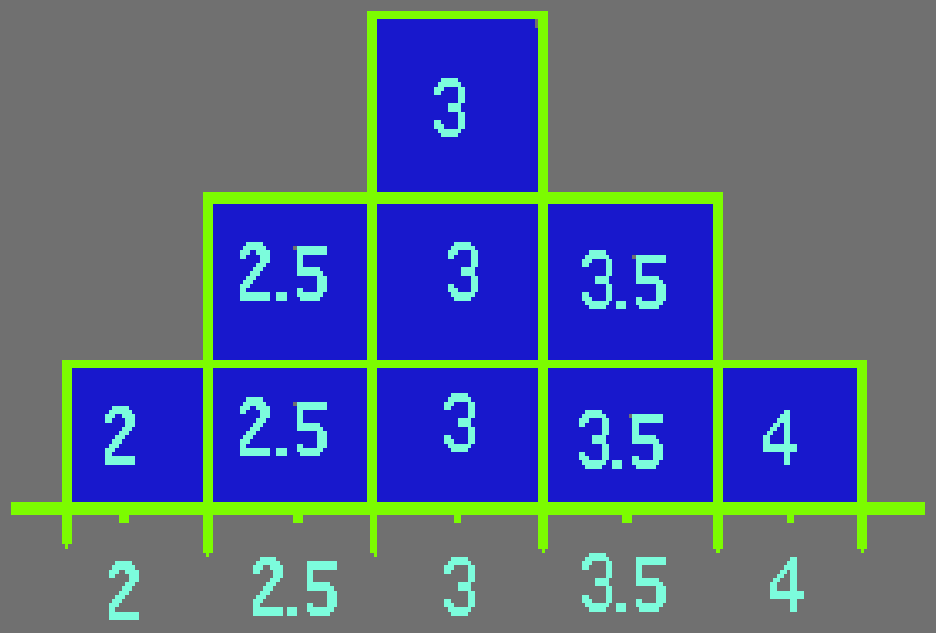
Even though the population is not normal, the sampling distribution will approximate a normal distribution.

The approximation becomes better as the sample size gets larger.

例 二

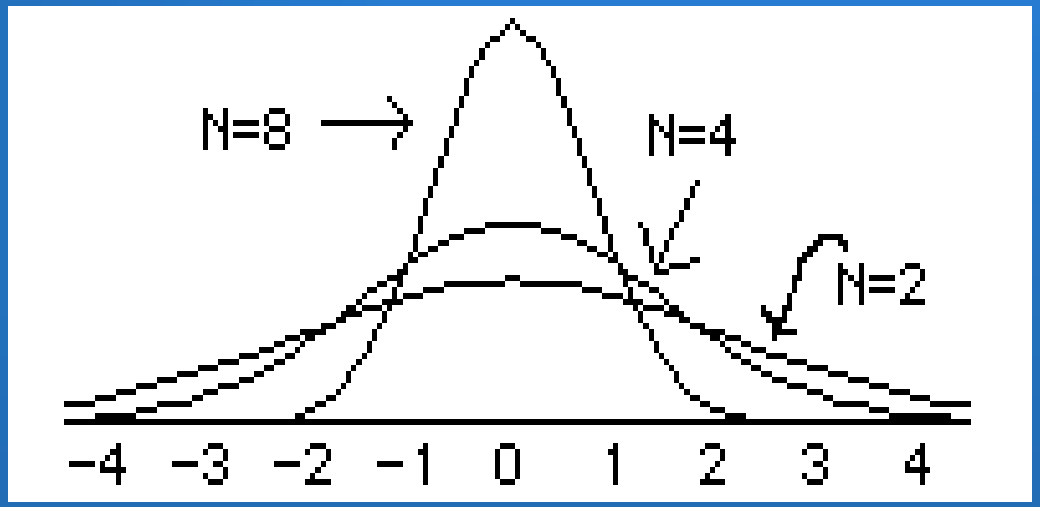
Possible items in Samples of n = 2

First Item	Second Item	Sample Mean
2	2	2
2	3	2.5
2	4	3
3	2	2.5
3	3	3
3	4	3.5
4	2	3
4	3	3.5
4	4	4



- 由以上模拟抽样试验可以看出，虽然原总体并非正态分布，但从中随机抽取样本，即使样本含量很小(n=2, n=4)，样本平均数的分布却趋向于正态分布形式。
- 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，$\overline{X}$ 的分布近似正态分布。



中心极限定理告诉我们：

- 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$\overline{X}$的分布是正态的。
- 若x的分布不很偏倚，在$n>20$时$\overline{X}$的分布就近似于正态分布了。

以抽样实验验证抽样分布：

例三：

设有一个N=4的有限总体，变数为2、3、3、4。根据 $\mu = \Sigma x / N$ 和 $\sigma^2 = \Sigma (x-\mu)^2 / N$ 求得该抽样总体的 μ、σ^2、σ 为：

$$\mu = 3,\ \sigma^2 = 1/2,\ \sigma = \sqrt{1/2} = 0.707$$

上述总体，如果按$n = 2$抽取样本，共可得$4^2=16$个样本；如果n为4，则一共可抽得$4^4=256$个样本。分别求这些样本的平均数 $\bar{x}$ ，其次数分布如表4-1所示。

表4-1 N=4, n=2 和 n=4 时的次数分布

$N^n=4^2=16$				$N^n=4^4=256$			
$\bar{x}$	f	$f\bar{x}$	$f\bar{x}^2$	$\bar{x}$	f	$f\bar{x}$	$f\bar{x}^2$
2.0	1	2.0	4.00	2.00	1	2.00	4.0000
2.5	4	10.0	25.00	2.25	8	18.00	40.5000
3.0	6	18.0	54.00	2.50	28	70.00	175.0000
3.5	4	14.0	49.00	2.75	56	154.00	423.5000
4.0	1	4.0	16.00	3.00	70	210.00	630.0000
				3.25	56	182.00	591.5000
				3.50	28	98.00	343.0000
				3.75	8	30.00	112.5000
				4.00	1	4.00	16.0000
Σ	16	48.0	148.00	Σ	256	768.00	2336.0000

根据表4-1，在n=2的试验中，样本平均数，抽样总体的平均数、方差与标准差分别为：

$$\mu_{\bar{x}} = \sum \frac{f_i \bar{x}_i}{N^n} = 48.0/16 = 3 = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{\sum f_i(\bar{x}_i - \mu_{\bar{x}})^2}{N^n} = \frac{\sum f_i \bar{x}_i^2 - (\sum f_i \bar{x}_i)^2 / N^n}{N^n} = \frac{148 - 48^2/16}{16}$$

$$= \frac{4}{16} = \frac{1}{4} = (1/2)/2 = \frac{\sigma^2}{n}$$

$$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}^2} = \sqrt{1/4} = \sqrt{1/2}/\sqrt{2} = \frac{\sigma}{\sqrt{n}}$$

同理，可得n=4时：

$$\mu_{\bar{x}} = 768/256 = 3 = \mu$$

$$\sigma_{\bar{x}}^2 = 32/256 = 1/8 = (1/2)/4 = \sigma^2/n$$

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{8}} = \frac{\sqrt{1/2}}{\sqrt{4}} = \frac{\sigma}{\sqrt{n}}$$

这就验证$\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$的正确性

样本平均值抽样分布的形成过程可以概括成下图

总体N

↓

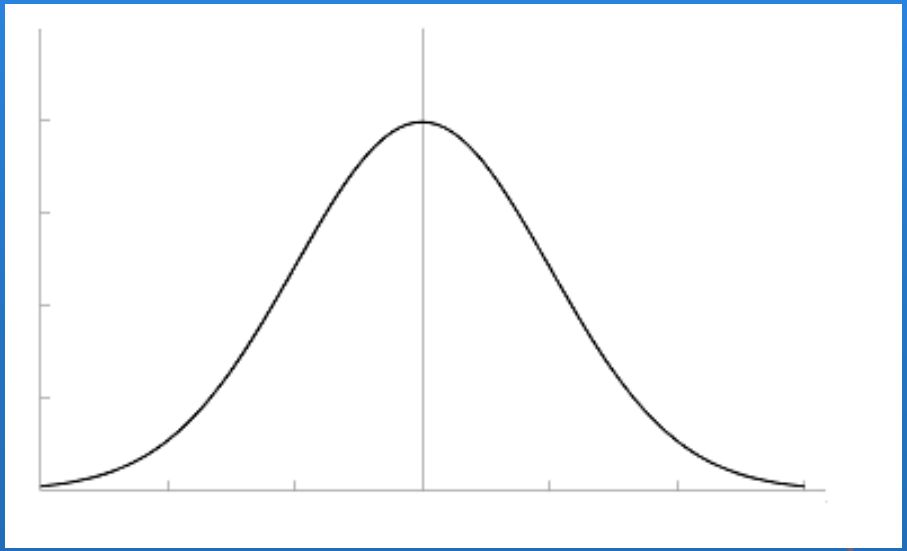
容量为n的所有样本

↓

计算出每一个样本的平均值

→

$\bar{x}$ 的抽样分布



抽样平均值的抽样分布与总体分布的关系

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

- 总体分布
  - 正态分布
    - 大样本
    - 小样本
    - 正态分布
  - 非正态分布
    - 大样本
      - 正态分布
    - 小样本
      - 近似正态分布

小 结

平均值分布的数学期望和方差

抽样方法	平均数	方　差	标准差
（1）从无限总体抽样和有限总体放回抽样	$E(\bar{x})=\bar{\bar{x}}=\mu$	$\sigma_{\bar{x}}^2=\dfrac{\sigma^2}{n}$	$\sigma_{\bar{x}}=\dfrac{\sigma}{\sqrt{n}}$ 抽样误差

★ $\sigma_{\bar{x}}=\dfrac{\sigma^2}{\sqrt{n}}$ 即平均数的抽样误差

4.2 从一个正态总体中抽样

4.2.1 样本平均数的分布

1. 总体σ已知时，样本平均数的分布—— 正态分布

2. 总体σ未知时，样本平均数的分布—— t 分布

4.2.2 样本方差的分布 – χ2分布

1. 总体σ已知时，样本平均数的分布—— 正态分布

- 根据中心极限定理可知，从 $N(\mu,\ \sigma^2)$的正态总体中独立随机地抽取含量为n的样本，则样本平均数分布服从 $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$

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

- 标准正态分布统计量，

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

平均数的标准误

- 标准误(平均数抽样总体的标准差) $\sigma_{\bar{x}} = \sigma / \sqrt{n}$ 的大小反映样本平均数 $\bar{x}$ 的抽样误差的大小，即精确性的高低。标准误大，说明各样本平均数 $\bar{x}$ 间差异程度大，样本平均数的精确性低。反之，$\sigma_{\bar{x}}$ 小，说明 $\bar{x}$ 间的差异程度小，样本平均数的精确性高。

- $\sigma_{\bar{x}}$ 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 是一常数，所以只有增大**样本含量**才能降低样本平均数的抽样误差。

2.总体σ未知时，样本平均数的分布- t分布

- 在实际工作中，总体标准差σ 往往是未知的，因而无法求得 $\sigma_{\bar{x}}$ 。
- 此时可用样本标准差 s 估计 σ 。

  以 $s/\sqrt{n}$ 估计 $\sigma_{\bar{x}}$ 。记$s/\sqrt{n}$为$S_{\bar{x}}$，称作**样本标准误或均数标准误。**
- $S_{\bar{x}}$ 是平均数抽样误差 $\sigma_{\bar{x}}$ 的估计值。

- 若样本中各观测值为 x_1，x_2，…，x_n，则

$$u=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\sim N(0,1) \longrightarrow t_{n-1}=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}=\frac{\bar{x}-\mu}{S_{\bar{x}}}$$

$$S_{\bar{x}}=\frac{S}{\sqrt{n}}=\sqrt{\frac{\sum(x-\bar{x})^2}{n(n-1)}}=\sqrt{\frac{\sum x^2-(\sum x)^2/n}{n(n-1)}}$$

- 样本标准差 s 与样本标准误 $S_{\bar{x}}$ 是既有联系又有区别的两个统计量
- 样本标准误公式 $S_{\bar{x}}=\frac{S}{\sqrt{n}}$ 表明了二者的联系。
- 区别在于：

1. 样本标准差 s 是反映样本中各观测值 $x_1, x_2, \ldots, x_n$ 变异程度大小的一个指标，它的大小说明 $\bar{x}$ 对该样本代表性的强弱。

2. 样本标准误 $S_{\bar{x}}$ 是样本平均数 $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ 的标准差，它是抽样误差的估计值，其大小说明了样本间变异程度的大小及精确性的高低。

- 对于大样本，常将样本标准差S与样本平均数 $\bar{x}$ 配合使用，记为 $\bar{x} \pm s$ ，用以说明所考察性状或指标的优良性与稳定性。
- 对于小样本，常将样本标准误 $S_{\bar{x}}$与样本平均数 $\bar{x}$ 配合使用，记为$\bar{x} \pm S_{\bar{x}}$（$\bar{x} \pm SE$ ），用以表示所考察性状或指标的优良性与抽样误差的大小。

- **t分布的概率分布密度函数如下：**

$$f(t)=\frac{1}{\sqrt{\pi df}}\frac{\Gamma[(df+1)/2]}{\Gamma(df/2)}(1+\frac{t^2}{df})^{-\frac{df+1}{2}}$$

式中，*t* 的取值范围是（-∞，+∞）；

***df*=*n*-1**为自由度。

- **t 分布的平均数和标准差为：**

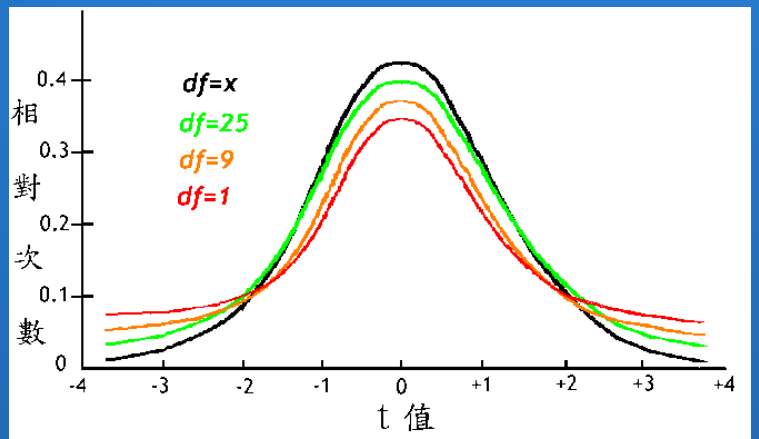
μ_t**＝0 (*df*>1)**，$\sigma_t=\sqrt{df/(df-2)}$ **(*df*>2）**

*t*分布密度曲线如图所示，其特点是：

1、*t*分布受自由度的制约，每一个自由度都有一条*t*分布密度曲线。

2、*t*分布密度曲线以纵轴为对称轴，左右对称，且在*t*＝0时，分布密度函数取得最大值。

3、与标准正态分布曲线相比，*t*分布曲线顶部略低，两尾部稍高而平。*df*越小这种趋势越明显。

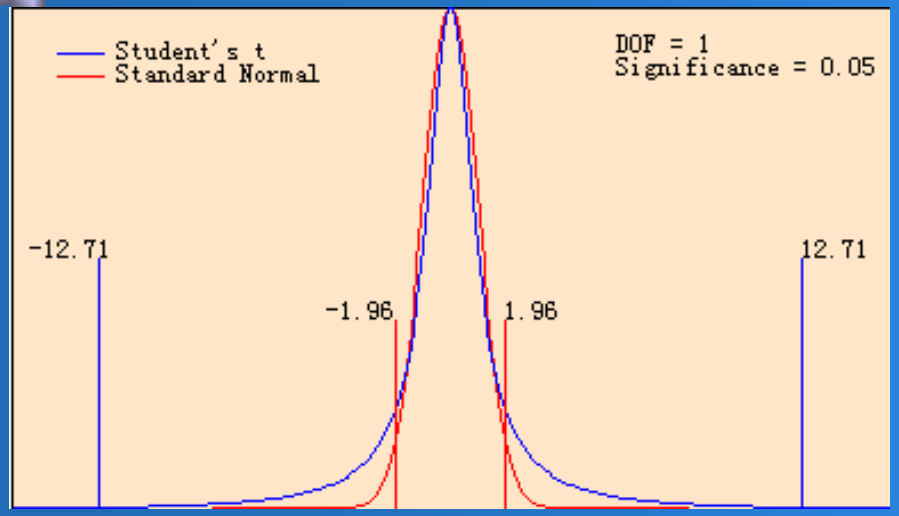


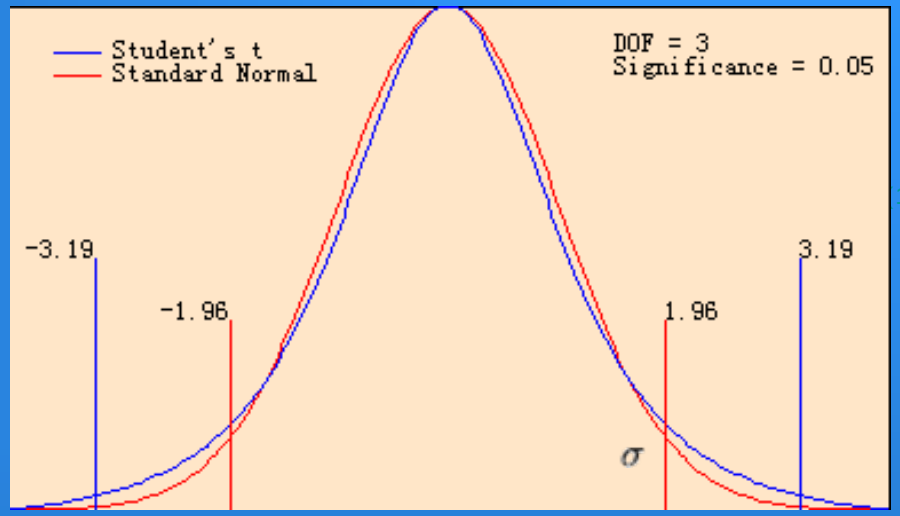
當df為1,9,25時的 t 分配曲線和常態分配之比較

- df越大，t分布越趋近于标准正态分布。
- 当n >30时，t分布与标准正态分布的区别很小；

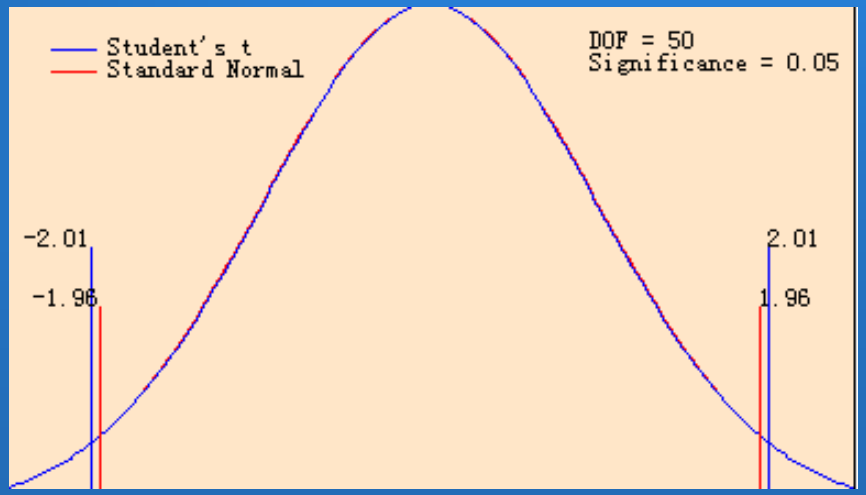
  n >100时，t分布基本与标准正态分布相同；

  $n \to \infty$时，t 分布与标准正态分布完全一致。


Student's t
Standard Normal
DOF = 1
Significance = 0.05
-12.71
-1.96
1.96
12.71


Student's t
Standard Normal
DOF = 3
Significance = 0.05
-3.19
-1.96
1.96
3.19
σ

(1)


Student's t
Standard Normal
DOF = 50
Significance = 0.05
-2.01
-1.96
2.01
1.96

(2)

- t分布的概率分布函数为：

$$F_{t(df)} = P(t < t_1) = \int_{-\infty}^{t_1} f(t)dt$$

1. t在区间（t_1，+∞）取值的概率-右尾概率为1-$F_{t(df)}$。由于t分布左右对称，t在区间（-∞，-t_1）取值的概率也为1-$F_{t(df)}$。

2. 于是t分布曲线下由-∞到-t_1和由t_1到+∞两个相等的概率之和—两尾概率为2(1-$F_{t(df)}$)。对于不同自由度下t分布的两尾概率及其对应的临界t值已编制成附表4，即t分布的临界值表。

- 例如，当df=15时，查附表4得两尾概率等于0.05的临界t值为2.131，其意义是：

$$P(-\infty<t<-2.131)= P(2.131<t<+\infty) =0.025$$

$$P(-\infty<t<-2.131)+ (2.131<t<+\infty) =0.05$$

- 由附表4可知：

当df一定时，
- 概率α越大，临界t值越小；
- 概率α越小，临界t值越大 。

当 α一定时，随着df 的增加，临界t 值在减小；

当df=∞时，临界t值与标准正态分布的临界u值相等。

4.2.2 样本方差的分布 – χ2分布

- 条件：

  从一个正态总体中抽样所得到的样本方差的分布

总体

$X \sim N(\mu, \sigma^2)$

抽样

样本

n, S^2

则 $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$

- 根据中心极限定理，从一个正态总体（ μ， σ² ）中抽取一个随机变量X，则统计量(x- μ)/ σ 也服从正态分布；
- 我们把这个统计量定义为 $U=\frac{x-\mu}{\sigma}$
- 再定义另一个统计量 $V=U^2=\left(\frac{x-\mu}{\sigma}\right)^2$
- 设抽得n个相互独立的随机变量x_i ,则各相应的V_i之和被称为χ^2，即 $$\chi^2=\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

这个统计量服从 ***df*=n-1** 的χ^2分布。

- 当 μ 未知时，用 $\overline{X}$ 代表 μ 来计算统计量

$$\chi^2=\sum\left(\frac{x_i-\mu}{\sigma}\right)^2=\frac{\sum\left(x_i-\overline{x}\right)^2}{\sigma^2}$$

由 $S^2=\dfrac{\sum\left(x_i-\overline{x}\right)^2}{n-1}$

$$\sum\left(x_i-\overline{x}\right)^2=(n-1)S^2$$

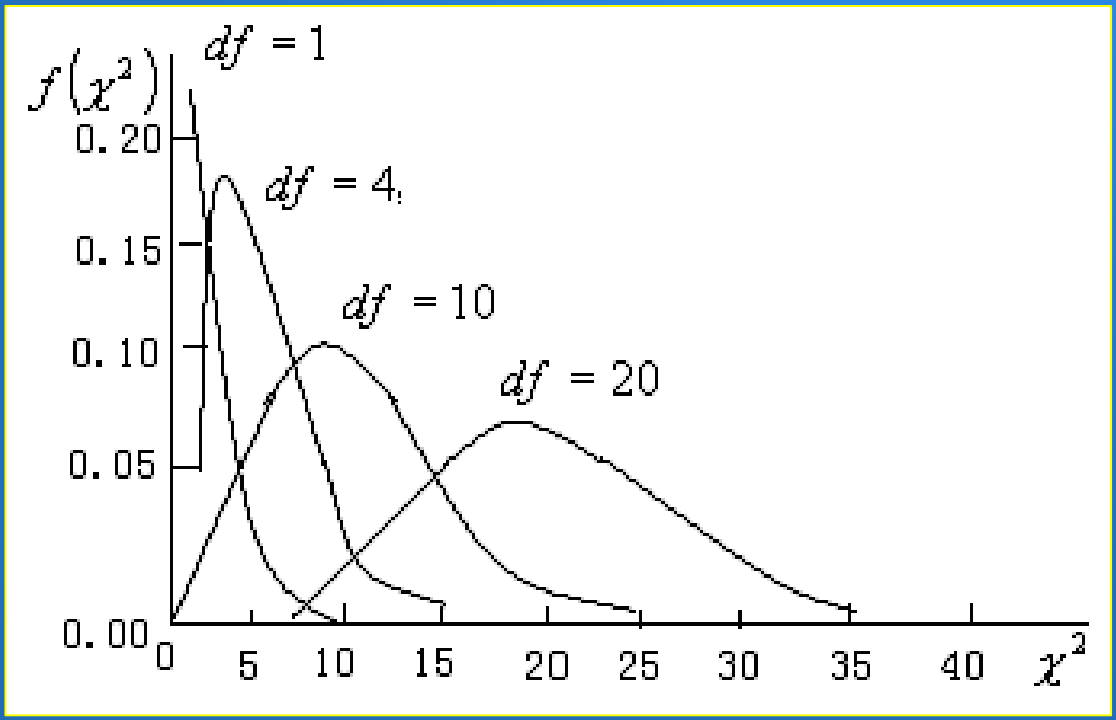
$$\therefore \chi^2=\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2=\frac{\sum\left(x_i-\mu\right)^2}{\sigma^2}=\frac{(n-1)\ S^2}{\sigma^2},df=n-1$$

随机样本$x_1, x_2, x_3,\ldots\ldots x_n$的统计量 $\dfrac{(n-1)S^2}{\sigma^2}$ 的分布是 **df=n-1的χ^2**分布。

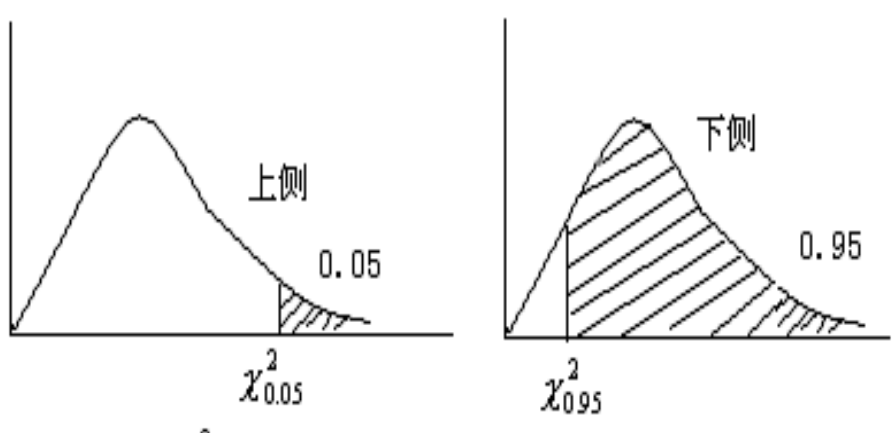
我们把这个分布就称为样本方差**s^2**的分布。

s^2分布的特征数为：

$$\mu_{s^2} = \sigma^2, \sigma_{s^2} = \sigma^2 \sqrt{\frac{2}{df}}$$



- 附表6是χ^2分布的上侧临界值表，例如：df=9,α=0.05 χ^2=16.919，或写为 P($\chi^2 \geq 16.919$)=0.05



χ^2分布的上侧、下侧分位点

- 若查下侧分位数，只要查出1- α的分位数即可。例如查df=9,α=0.05下侧分位数，只要查1-0.05=0.95得$\chi_{0.95}^2$ =3.33

4.3 从两个正态总体中抽取的样本统计量的分布

4.3.1两个样本平均数差的分布

1.标准差σ_i已知时，两个平均数差($\bar{x}_1 - \bar{x}_2$)的分布

2. σi未知但相等时，两个平均数差的分布

4.3.2 两个样本方差比的分布-F分布

两个样本平均数之差的抽样分布

总体 (μ_1, σ_1^2) —抽样→ 样本 x1, (n1)

总体 μ_2, σ_2^2 —抽样→ 样本 x2, (n2)

$$X_1 - N(\mu_1, \sigma_1^2),\ X_2 - N(\mu_2, \sigma_2^2),$$

则 $(\bar{x}_1 - \bar{x}_2) \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

$\bar{x}_1 - \bar{x}_2 = A$ —估计→ $\mu_1 - \mu_2 = ?$

4.3.1 两个样本平均数差的分布

1.标准差σi已知时，两个平均数差($\bar{x}_1-\bar{x}_2$)的分布-正态分布-

$$\mu_{(\bar{x}_1-\bar{x}_2)} = E\left(\bar{x}_1-\bar{x}_2\right)=E(\bar{x}_1)-E(\bar{x}_2)=\mu_1-\mu_2$$

$$Var\left(\bar{x}_1-\bar{x}_2\right)=Var\left(\bar{x}_1\right)+Var\left(-\bar{x}_2\right)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

即 $\bar{x}_1-\bar{x}_2$ 的分布服从正态分布 $N\left[(\mu_1-\mu_2),\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)\right]$

标准化得：

$$u=\frac{\left(\bar{x}_1-\bar{x}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

~ N（0，1）分布

2. σ_i未知但相等时，两个平均数差（$\bar{x}_1-\bar{x}_2$）的分布-t分布

当σ_1 ,σ_2未知但相等时，可用s_1^2和s_2^2代替之，得到统计量

$$t_{n_1+n_2-2}=\frac{\left(\bar{x}_1-\bar{x}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{\left(n_1-1\right)S_1^2+\left(n_2-1\right)S_2^2}{\left(n_1-1\right)+\left(n_2-1\right)}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

当$n_1=n_2$时，上式简化为：

$$t_{2n-2}=\frac{\left(\bar{x}_1-\bar{x}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{S_1^2+S_2^2}{n}}}$$

4.3.2 两个样本方差比的分布-F分布

- 方差比F的定义是：
  两个独立的具df_1, df_2服从χ^2分布的随机变量χ_1^2, χ_2^2各除以对应的自由度之比，即：

$$F=\frac{\chi_1^2}{df_1}\Big/\frac{\chi_2^2}{df_2}$$

已知 $\dfrac{\sum\left(x_i-\bar{x}\right)^2}{\sigma^2}=\dfrac{(n-1)S^2}{\sigma^2}$ 是具有$df=n-1$的χ^2分布

- $$F=\frac{(n_1-1)S_1^2}{\sigma_1^2(n_1-1)}\Big/\frac{(n_2-1)S_2^2}{\sigma_2^2(n_2-1)}=\frac{\dfrac{S_1^2}{\sigma_1^2}}{\dfrac{S_2^2}{\sigma_2^2}}$$ 当$\sigma_1^2=\sigma_2^2$时，$F=\dfrac{S_1^2}{S_2^2}$

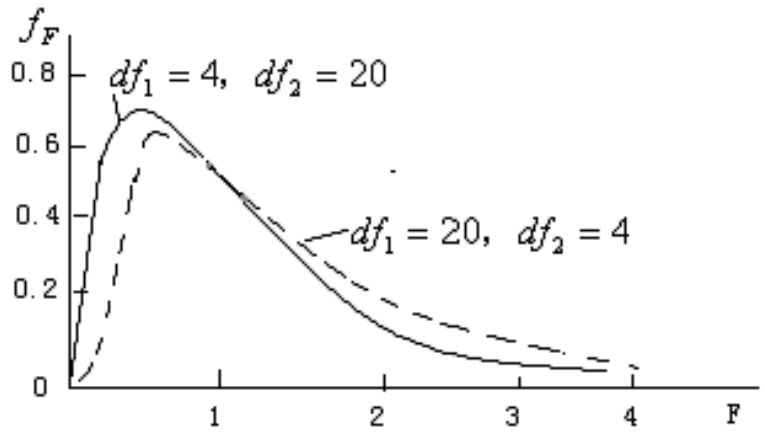
这个统计量服从F分布。它由一对自由度df_1,df_2决定。

F分布的特征数为：

$$\mu_F = \frac{df_2}{df_2 - 2} \quad , \quad \sigma_F^2 = \frac{2df_2^2\left(df_1 + df_2 - 2\right)}{df_1\left(df_2 - 2\right)^2\left(df_2 - 4\right)} \qquad df_2 \rangle 2$$

- 附表7是F分布的临界值（上侧分位数）表。
  查得 $df_1=20, df_2=4$ $\alpha=0.01$，$F_{0.01}=14.02$，
  为了查表方便，我们计算F值时，总是取F值大于1，
  不必查下侧临界值。

$$F = \frac{S_{大}^2}{S_{小}^2}$$



不同自由度下的F分布

注：$P(F \rangle 4.431) = 0.01$，$df_1 = 4$，$df_2 = 20$，$\alpha = 0.01$，$F = 4.431$