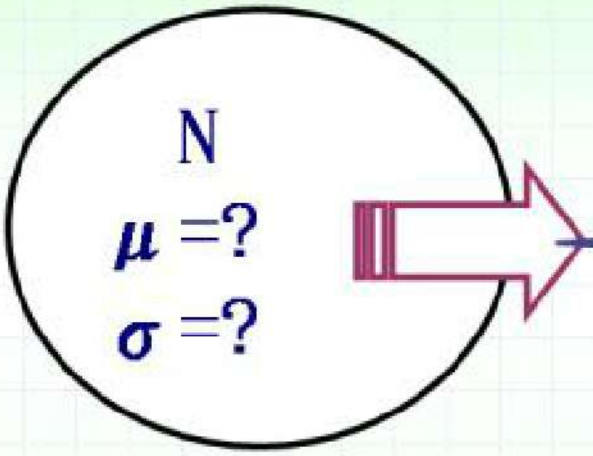
第四章

抽样分布与参数估计

概 述



$$\begin{cases} n_1 = 110 & n_2 = 100 \\ \overline{X}_1 = 121cm & \overline{X}_2 = 120cm \\ S_1 = 4.72cm & S_2 = 4.75cm \end{cases}$$

1995年某市7岁
男童身高总体

样本

抽样的目的……统计推断

统计推断： 由样本信息推断总体特征。

样本信息 → 样本统计指标（统计量）

总体特征 → 总体统计指标（参数）

统计推断
- 参数估计
- 假设检验

抽样分布:

从同一总体中,随机抽取相同含量的样本,由重复抽取的每一样本均可计算获得一个样本统计量,样本统计量所有可能取值的分布就是抽样分布.

均数的抽样分布:

从同一总体中,随机抽取相同含量的样本,由重复抽取的每一样本均可计算获得一个样本均数值,样本均数的分布称为均数的抽样分布.

统计推断过程中应考虑的问题:

1、上述两样本均数互不相等，也不一定等于总体均数，为什么?

2、上述两样本均数中，哪个与总体均数更接近，估计总体均数可靠?

3、用样本推断总体时，用何种方法，即样本统计量的抽样分布有何规律?

4、抽样误差的大小及计算?

第一节

抽样分布与标准误

一、样本均数的抽样分布与标准误

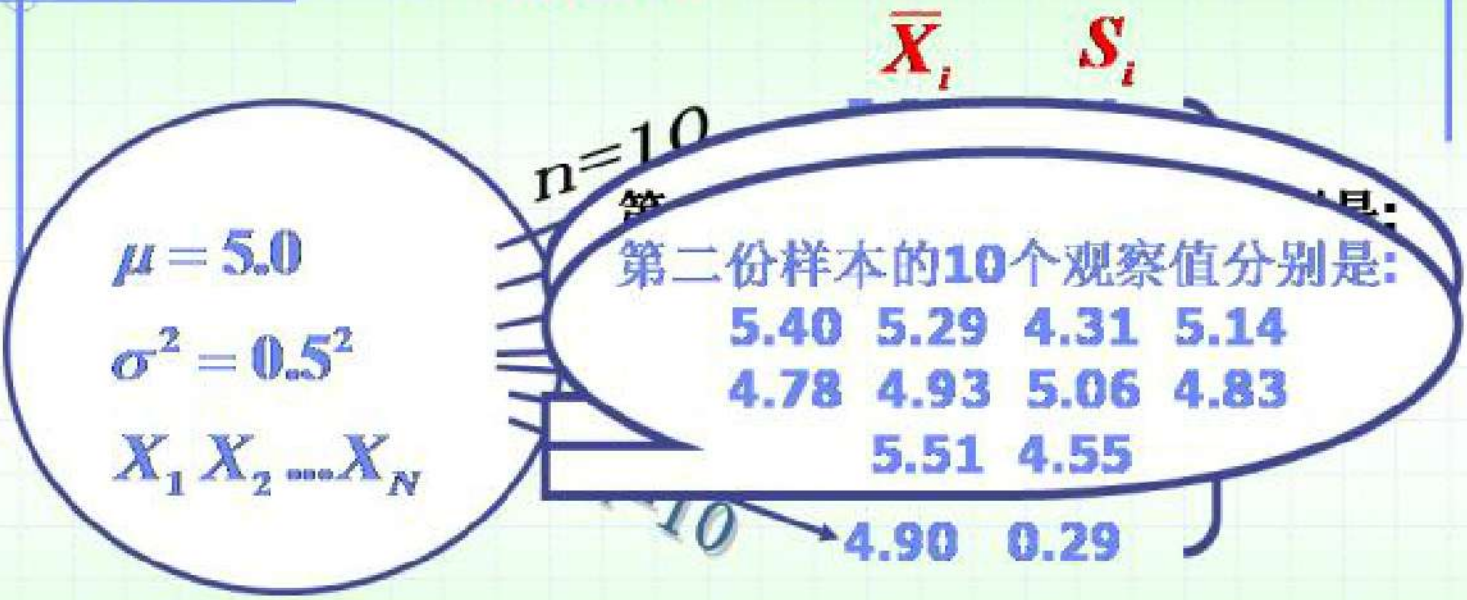
均数的抽样分布有什么规律？

什么是均数的抽样误差？

标准误的计算及应用？

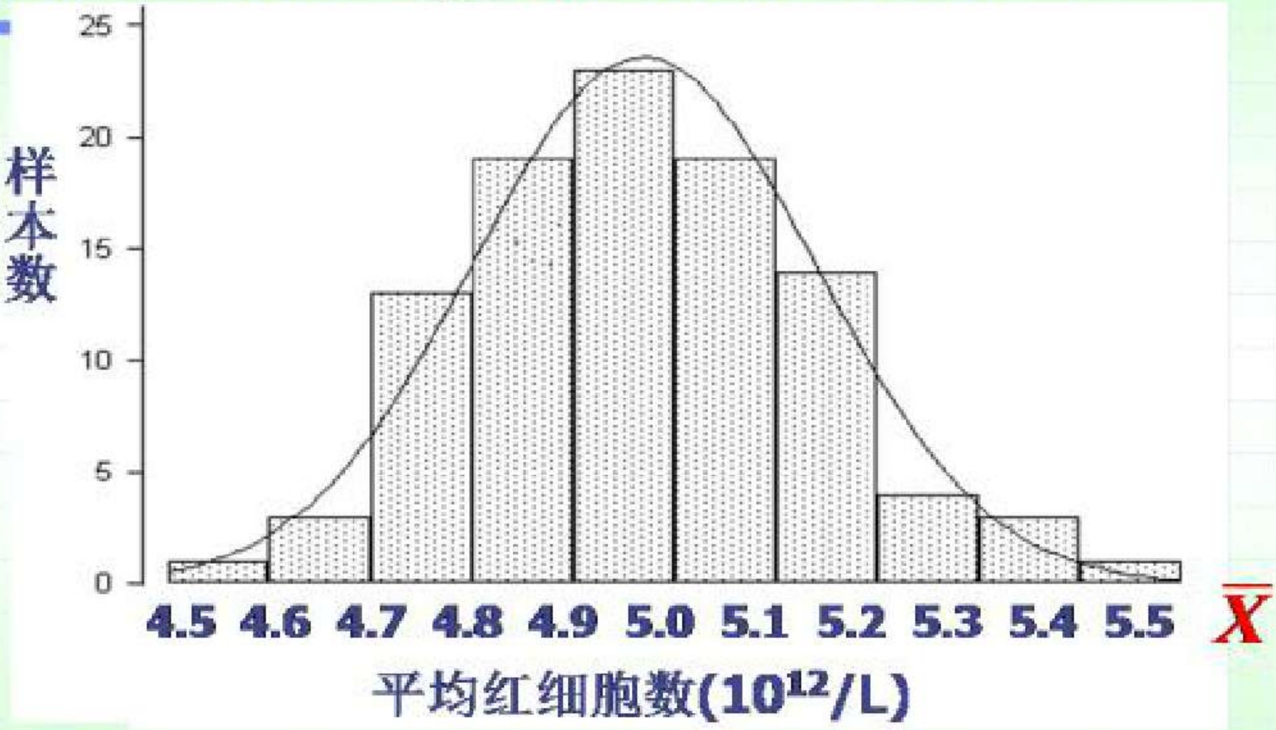
请看下面的模拟抽样

假定正常成年男子红细胞计数服从$N(5.00,0.50^2)$的正态分布总体，从该总体中重复进行100次抽样，每个样本含量为10，抽样过程见下面（模拟抽样1）：

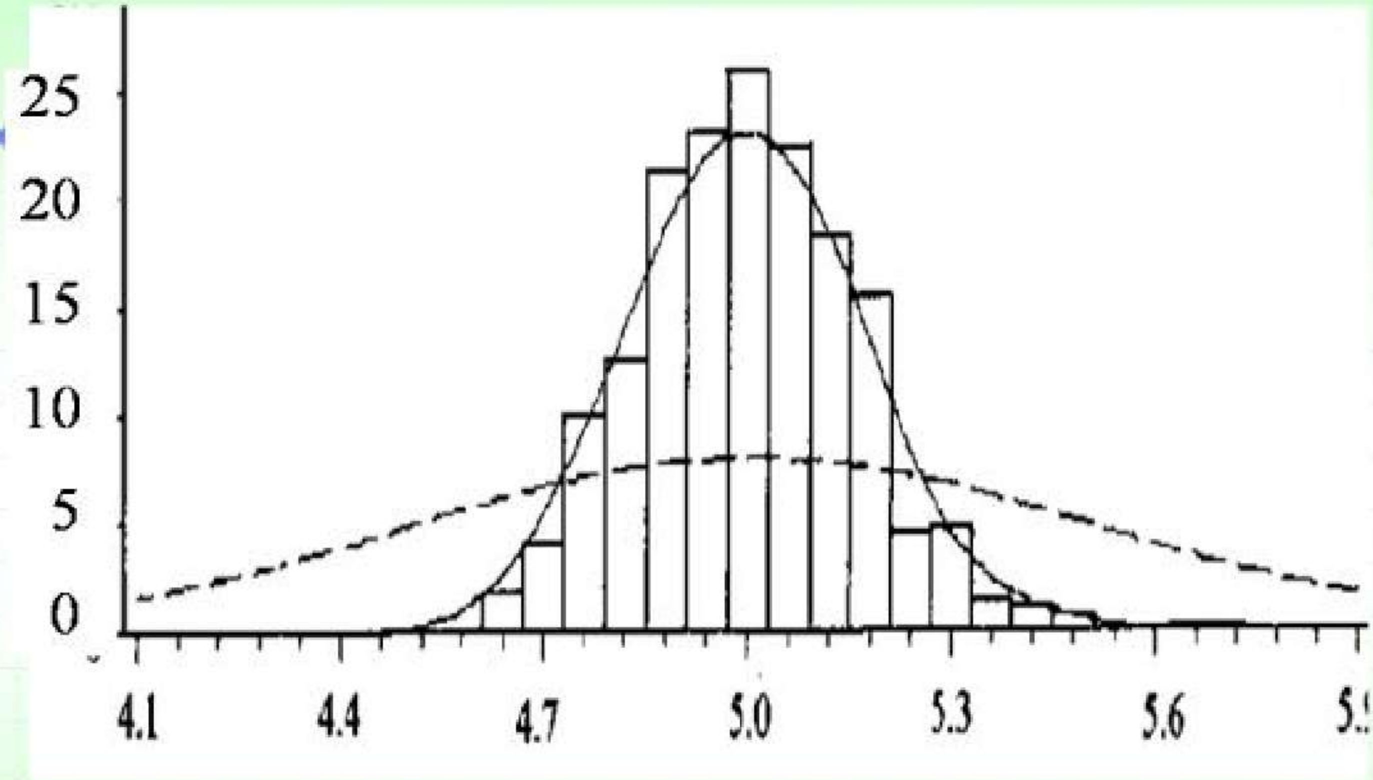


正常成年男子红细胞数$N(5.00, 0.50^2)10^{12}/L$的抽样示意图

将此100个样本均数看成**新变量值**，则这100个样本均数构成一**新分布**，绘制直方图。



从正态分布总体$N(5.00, 0.5^2)$中随机抽样所得样本均数分布



新分布与原总体分布（母分布）的关系

总结：这100个样本均数的分布近似正态分布，其均数和标准差分别为

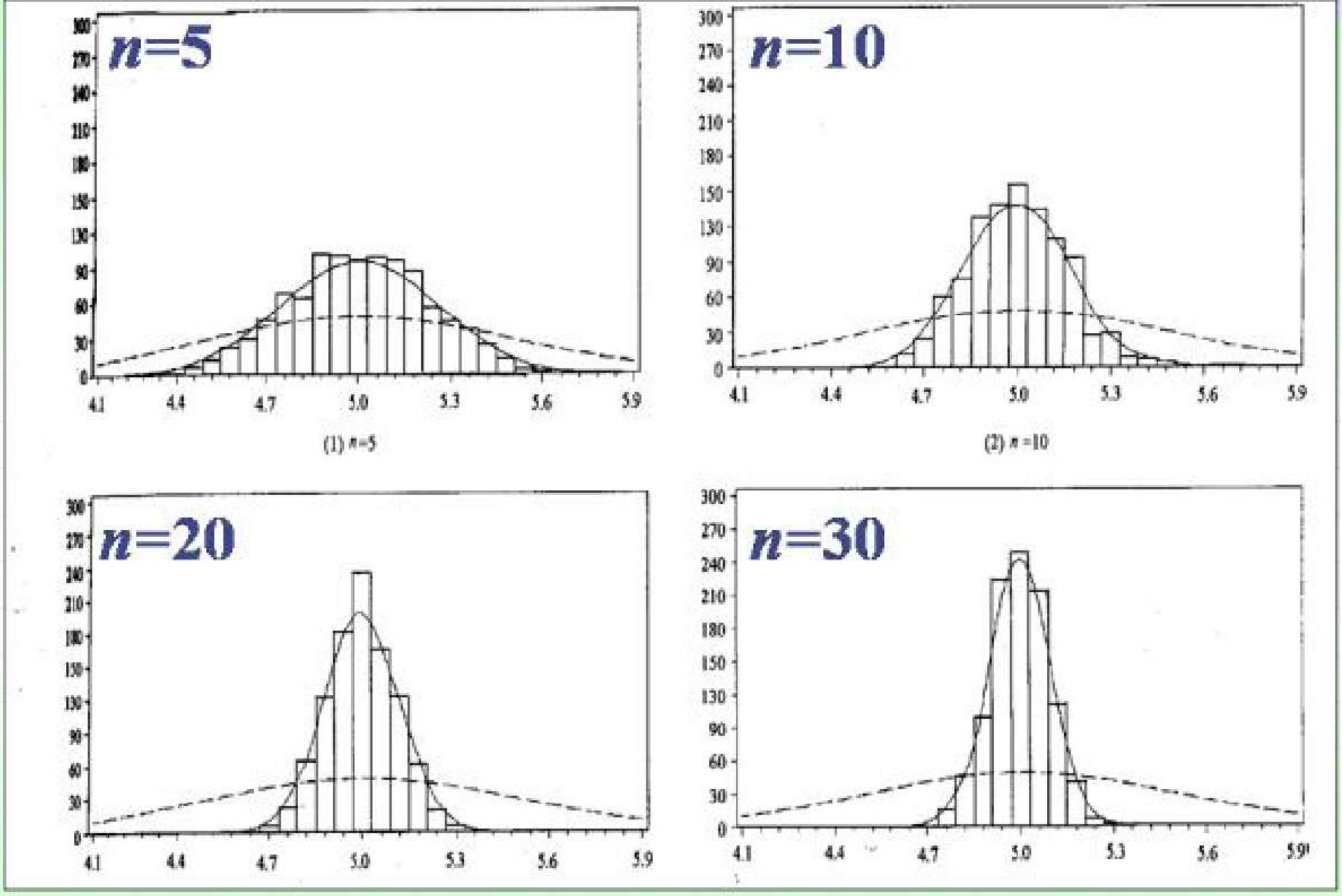
100个样本均数的均数 $\bar{\bar{X}}$

$$\bar{\bar{X}}=\frac{\bar{X}_1+\bar{X}_2+...+\bar{X}_{100}}{100}=5.09\approx\mu=5.0$$

100个样本均数的标准差 $s_{\bar{X}}$

$$s_{\bar{X}}=\sqrt{\frac{\sum_{i=1}^{100}(\bar{X}_i-\bar{\bar{X}})^2}{100-1}}=0.157\approx\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{10}}=0.16$$

模拟抽样2： 假定正常成年男子的红细胞计数服从正态分布N（$5.00, 0.50^2$）$\times 10^{12}$/L，从中做随机抽样实验，每次从这一总体中随机抽取1000份样本，每份样本含量分别为5、10、20、30，分别计算每份样本均数，对不同样本含量的抽样实验的1000个样本均数作直方图并拟合分布曲线。如图



从正态分布N(5.00,0.50²)总体中抽样实验结果

样本均数（正态总体中抽样）的抽样分布具有如下特点:

① $\bar{X} \neq \mu$，各样本均数 $\bar{X}$ 未必等于总体均数

各样本均数间也互不相等。

② 样本均数的分布以 $\bar{\bar{X}} \approx \mu$ 为中心，左右基本对称，近似正态。

③ 样本均数的变异程度 $\sigma_{\bar{X}}$ 较之原变量的变异程度 σ 大大减小，且 n 越大，$\sigma_{\bar{X}}$ 越小。

可证明：$\mu_{\bar{X}} = \mu_X$；$\sigma_{\bar{X}} < \sigma_X$。

1、均数的抽样误差

- 概念：由抽样造成的样本统计量之间（如 $\overline{X}$ 之间）以及样本统计量与总体参数间（如 $\overline{X}$ 与 μ 间）的差异。
- 原因：
  - 总体中个体差异大小即 σ 的大小
  - 抽取的样本含量大小即 n 的大小
  - 抽样方法
- 大小：标准误反映

2、标准误

- 概念：样本均数的标准差，反映均数抽样误差大小的指标 。
- 符号：$\sigma_{\bar{X}}$（理论值），$S_{\bar{X}}$（估计值）
- 公式：

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

- 应用：实际工作中通常只能获得 $S_{\bar{X}}$

例：以第二号样本为例

已知： $\overline{X}=5.03, S=0.52, n=10$

则：

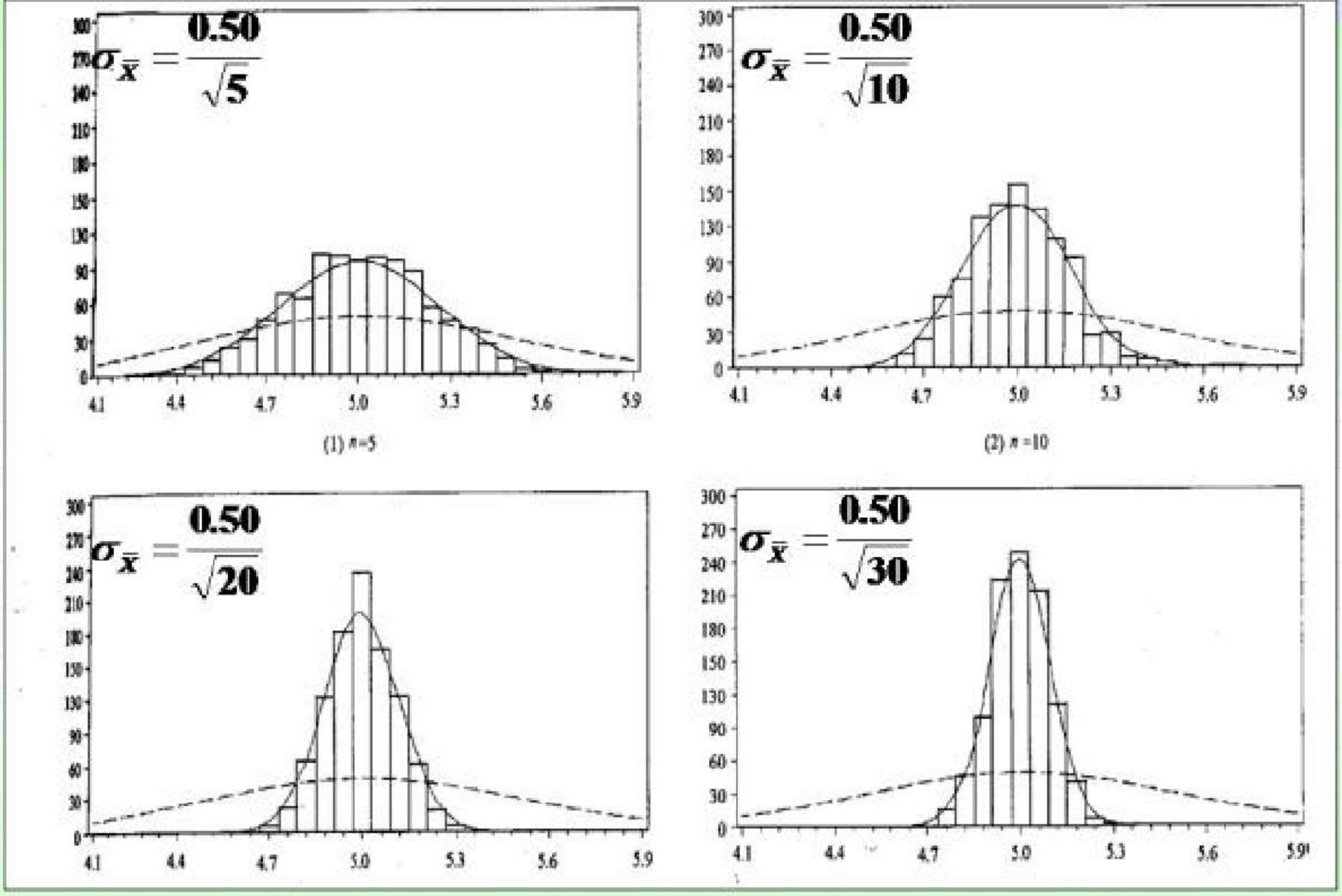
$$S_{\overline{X}}=\frac{S}{\sqrt{n}}=\frac{0.52}{\sqrt{10}}=0.1644(10^{12}/L)$$

$$\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}=\frac{0.50}{\sqrt{10}}=0.16(10^{12}/L)$$

可见： $S_{\overline{X}}\approx\sigma_{\overline{X}}$ ，可用 $S_{\overline{X}}$ 作为 $\sigma_{\overline{X}}$ 的估计值。

$\sigma_{\bar{x}}$与$S_{\bar{x}}$的比较

- $\sigma_{\bar{x}}$ 是一个可变的常数，即n一定时，$\sigma_{\bar{x}}$ 是个常数；但n改变，$\sigma_{\bar{x}}$ 也改变，且$n\to$大，$\sigma_{\bar{x}}\to$小；$n\to\infty$，$\sigma_{\bar{x}}\to 0$。
- $S_{\bar{x}}$ 是个可变的数，即n一定时，样本不同，$S_{\bar{x}}$不同；n越小，$S_{\bar{x}}$的变异越大，$n\to$大，$S_{\bar{x}}$的变异逐渐减小，$n\to\infty$，$S_{\bar{x}}\to\sigma_{\bar{x}}$



从正态分布N(5.00,0.50²)总体中抽样实验结果

3、标准误的应用

- 反映抽样误差的大小，标准误越小，说明抽样误差越小，样本均数与总体均数越接近，用样本均数估计总体均数越可靠，反之亦然。
- 用于确定总体均数的可信区间。

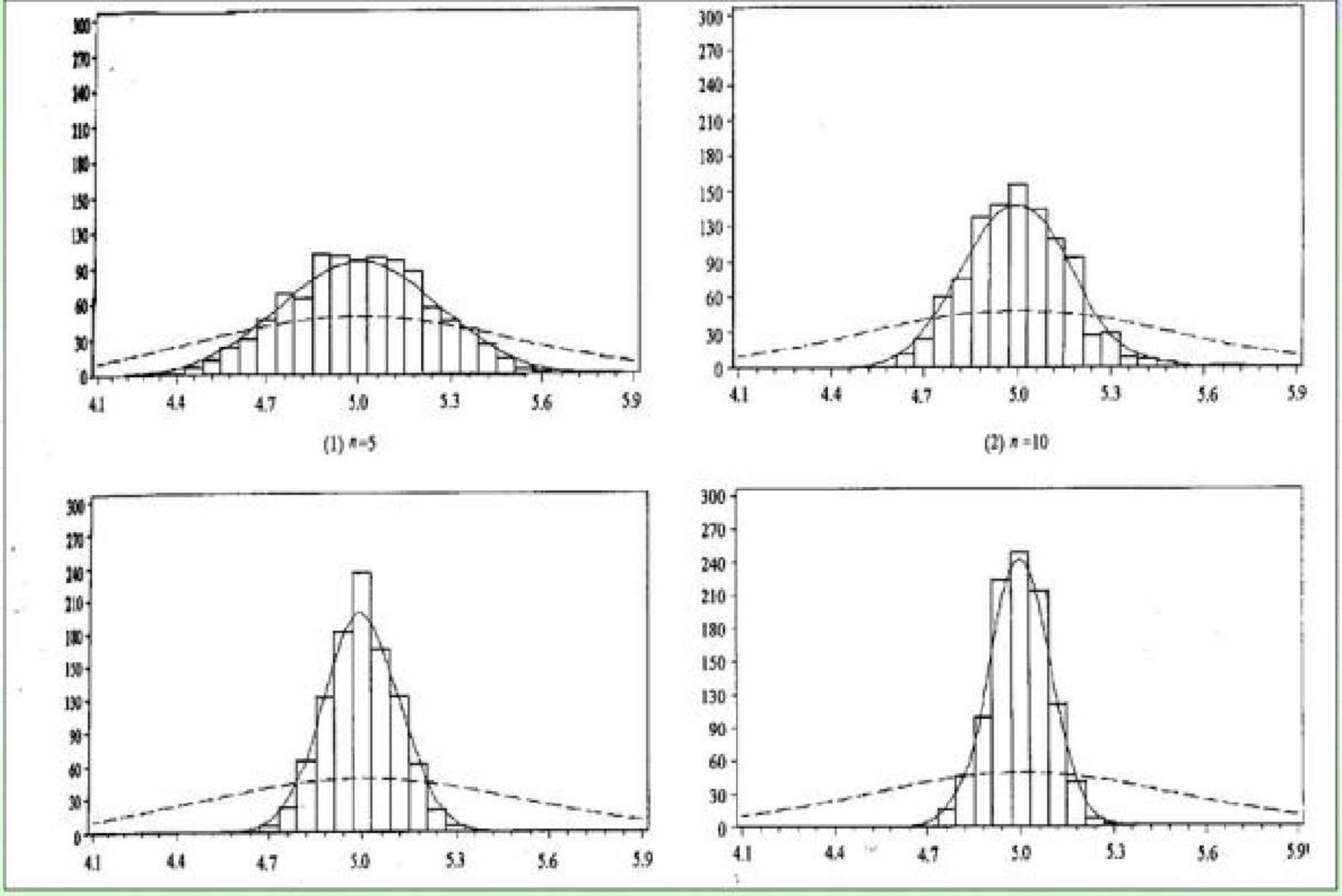
注意：降低抽样误差（即 $\downarrow S_{\bar{X}}$ ）的途径有：

①通过增加样本含量 n；

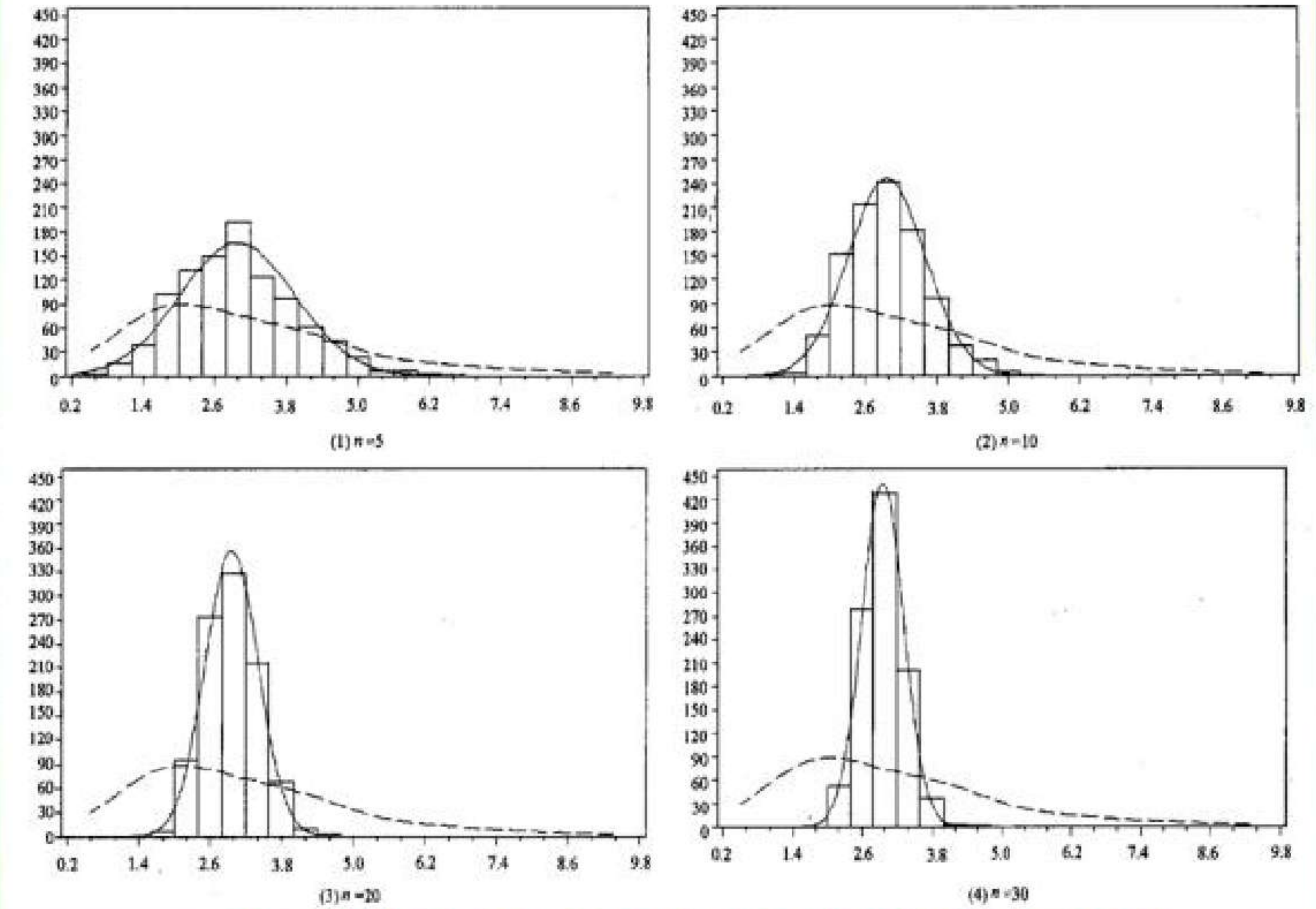
②通过设计减少 S。

4、抽样的中心极限定理

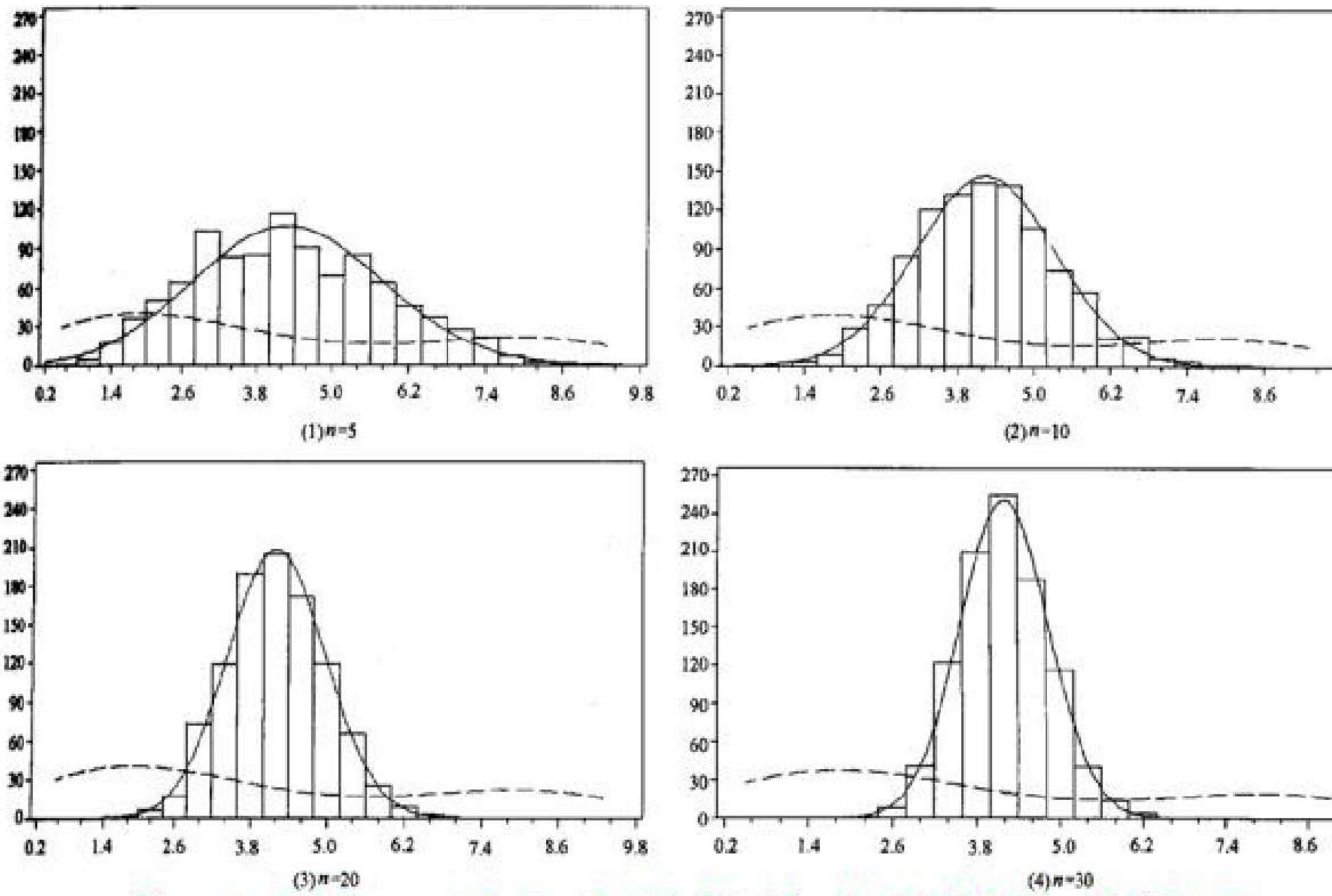
从正态分布、偏态分布总体中作抽样实验（4次），每次从这一总体中随机抽取1000份样本，第1、2、3、4次实验的每份样本含量分别为5、10、20、30，分别作直方图如下：



从正态分布总体中抽样实验结果



从正偏态分布总体中抽样实验结果



从不对称凹型分布的总体中抽样实验结果

总结：均数的抽样分布

- 若X_i服从正态分布，则$\overline{X}_i$亦服从正态分布。
- 若X_i不服从正态分布，则

  n小，$\overline{X}_i$为非正态分布 （p_{65}）

  n大，$\overline{X}_i$为正态分布
- 一般，只要$n>30$，可认为$\overline{X}_i$服从正态分布，即简记为$\overline{X} \sim N(\mu, \sigma_{\overline{x}}^2)$

二、两个样本均数间差值的抽样分布与标准误

将 $\bar{X}_1-\bar{X}_2$ 看成 $\bar{X}$，则两个样本均数间差值的抽样分布与两样本含量大小及两样本所来自的总体分布有关。

1、大样本（n_1和n_2都≥30）

* 两个样本均数间差值的抽样分布近似正态

两个样本均数间差值 $\bar{X}_1-\bar{X}_2$ 的均数为:

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

两个样本均数间差值 $\bar{X}_1-\bar{X}_2$的标准误为:

$$S_{\bar{X}_1-\bar{X}_2}=\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

2、小样本（n_1或n_2小于30）

如果$\bar{X}_1, \bar{X}_2$分别来自正态分布(μ_1, σ_1^2)，(μ_2, σ_2^2)

则两样本均数间差值的抽样分布近似正态

两个样本均数间差值$\bar{X}_1 - \bar{X}_2$的均数为：

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

两个样本均数间差值$\bar{X}_1 - \bar{X}_2$的标准误为：

1 $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$（$\sigma_1, \sigma_2$已知，少见）

σ_1, σ_2未知（最常见），但$\sigma_1^2=\sigma_2^2$时

$$S_{\bar{X}_1-\bar{X}_2}=\sqrt{S_C^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}=S_C\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$S_C=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}}$$

$$=\sqrt{\frac{\sum X_1^2-(\sum X_1)^2/n_1+\sum X_2^2-(\sum X_2)^2/n_2}{n_1+n_2-2}}$$

三、样本率的抽样分布与标准误

样本率的抽样分布与样本含量n的大小及样本率p值的大小有关。

1、样本率

样本率为 $p=\frac{X}{n}$

样本率p 的总体均数为 $\mu_p=\pi$

样本率p的总体方差为 $\sigma_p^2=\frac{\pi(1-\pi)}{n}$

样本率p的总体标准差（即率的标准误）为

$$\sigma_p=\sqrt{\frac{\pi(1-\pi)}{n}}$$

2、率的标准误

❶**概念**：样本率的标准差也称为率的标准误，可用来反映样本率的抽样误差大小。

❷**意义**：率的标准误越小，则率的抽样误差就越小，用样本率估计总体率越可靠；反之亦然。

❸**公式**：π已知，用理论值，否则，用估计值。

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}(\text{理论值}), S_p = \sqrt{\frac{p(1-p)}{n}}(\text{估计值})$$

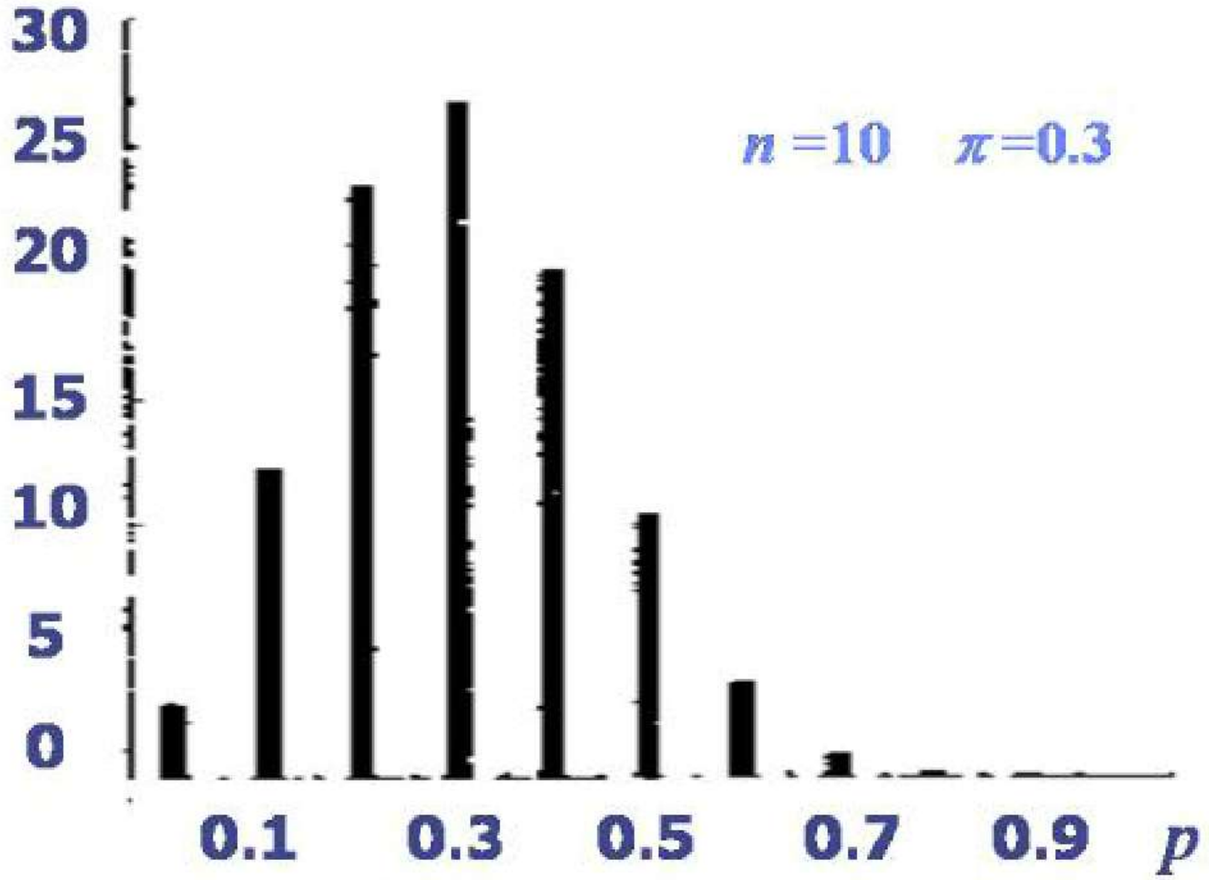
3、样本率的抽样分布

❶ n较小时，样本率的抽样分布呈离散性偏态分布。前进

❷ 当n足够大，样本率p值不接近于0与1，满足np和$n(1-p)$均≥ 5时，p的抽样分布逼近正态分布。即：

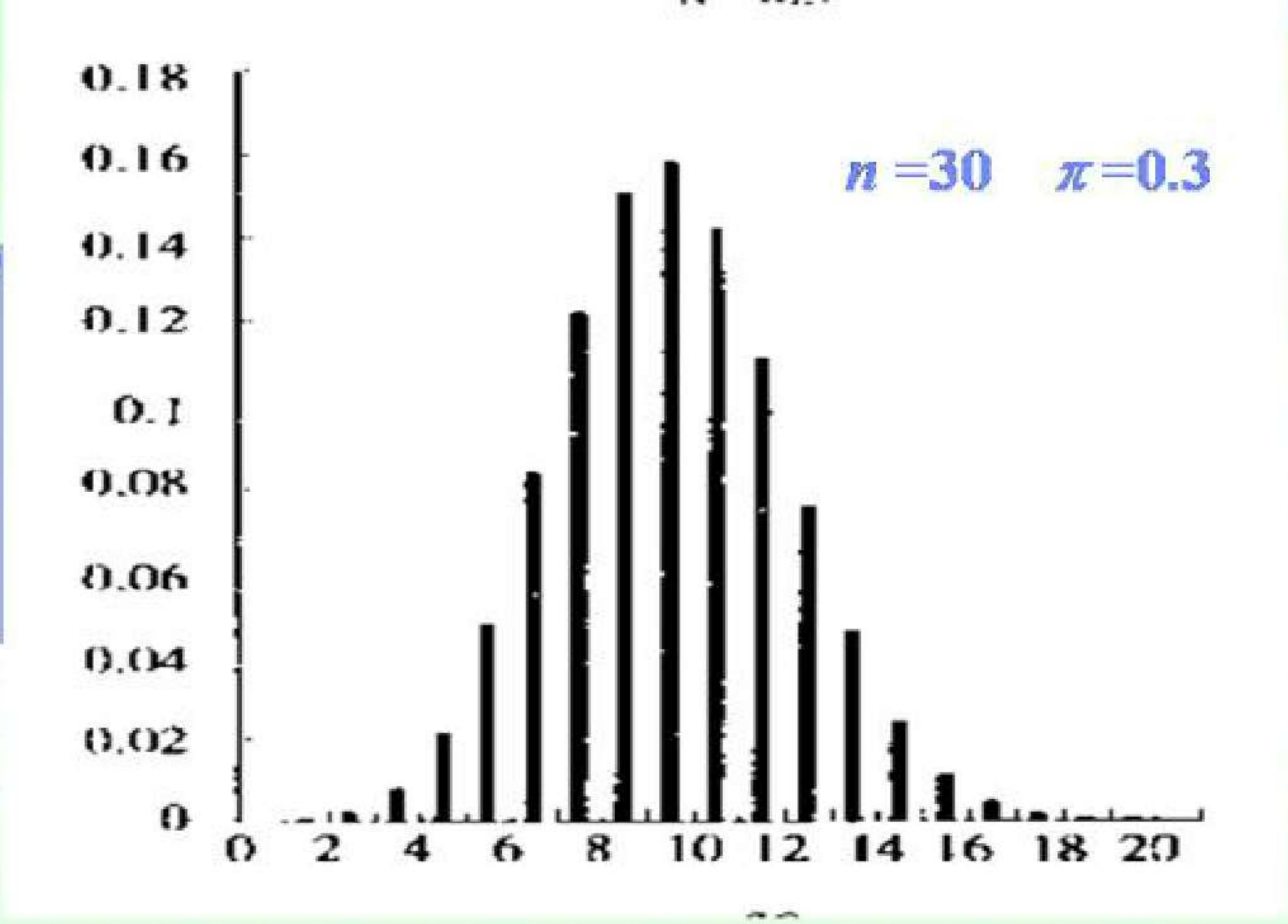
$$p \sim N(\pi, \sigma_p)$$

33张



从$\pi=0.3$的总体中抽出的100个样本率的分布

返回



从π=0.3的总体中抽出的100个样本率的分布

四、两个样本率间差值的抽样分布与标准误

两个样本率间差值的抽样分布与两个样本含量n_1、n_2的大小及两个样本率值p_1、p_2的大小有关。

1、两个样本率间差值 p_1-p_2

p_1-p_2 的总体均数为 $\mu_{p_1-p_2}=\pi_1-\pi_2$

p_1-p_2 的标准误为（即率的标准误）为

$$S_{p_1-p_2}=\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$p_c=\frac{X_1+X_2}{n_1+n_2}$$

2、两个样本率间差值 p_1-p_2的抽样分布

当n_1, n_2均足够大，p_1, p_2不接近0或1，满足$n_1p_1, n_1(1-p_1)$ 与$n_2p_2, n_2(1-p_2)$ 均≥ 5，两样本率间差值的抽样分布近似正态分布。

第二节

常见的几种抽样分布

t分布（t-distribution）

一、t 分布的概念

1．若某一随机变量 X 服从总体均数为 μ、总体标准差为σ 的正态分布$N(\mu, \sigma^2)$，则可通过 z 变换(即 $z=\frac{X-\mu}{\sigma}$)将一般正态分布转化为标准正态分布 $N(0,1^2)$，即 z 分布；

2．若样本均数$\bar{X}$服从总体均数为μ、总体标准差为$\sigma_{\bar{X}}$的正态分布$N(\mu,\ \sigma_{\bar{X}}^2)$，则通过同样方式的$z$变换($\frac{\bar{X}-\mu}{\sigma_{\bar{X}}}$)也可将其转换为标准正态分布$N(0,1^2)$，即$z$分布。

3. 实际工作中，由于 $\sigma_{\bar{X}}$未知，用 $S_{\bar{X}}$代替，则 $(\bar{X}-\mu)/S_{\bar{X}}$不再服从标准正态分布z分布，而服从t分布。

$$t=\frac{\bar{X}-\mu}{S_{\bar{X}}}=\frac{\bar{X}-\mu}{S/\sqrt{n}},\quad \nu=n-1$$

式中 ν 为自由度(degree of freedom, *df*)。

二、t 分布的图形与特征

1、图形

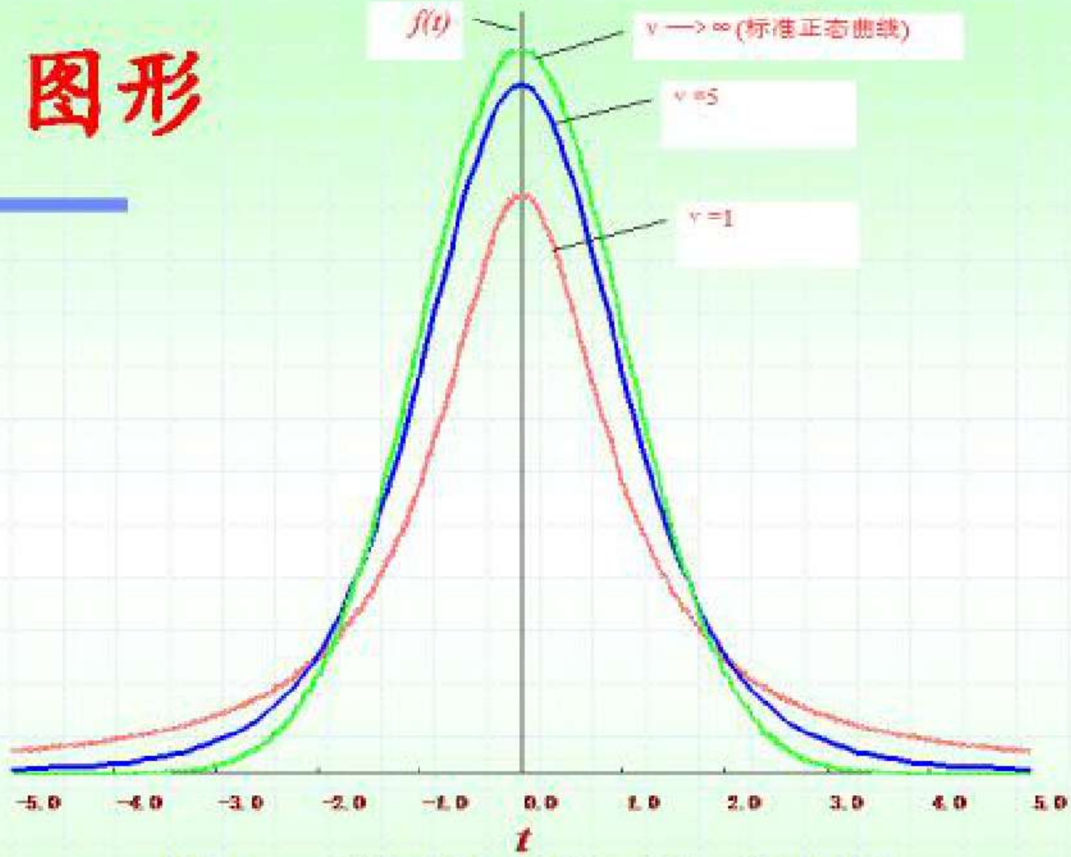


图4-3 不同自由度下的t分布图

$$f(t)=\frac{\Gamma\left[(\nu+1)/2\right]}{\sqrt{\pi\nu}\Gamma(\nu/2)}(1+\frac{t^2}{\nu})^{-\frac{(\nu+1)}{2}} \quad -\infty < t < \infty$$

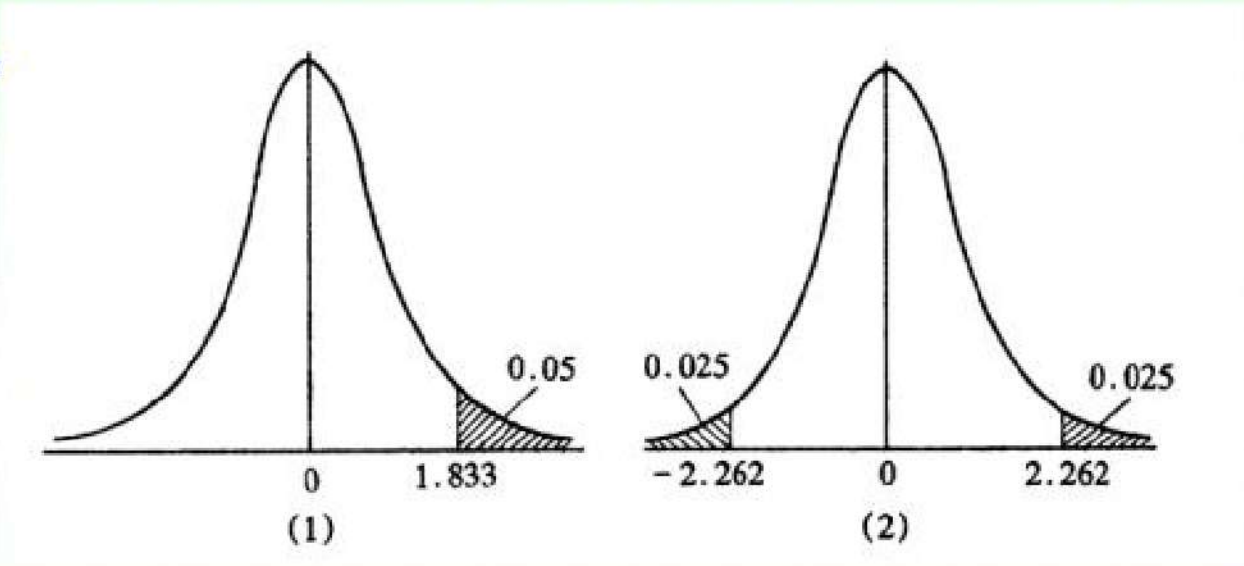
2、特征

① t 分布为单峰分布，曲线在 $t=0$ 处最高，并以 $t=0$ 为中心左右对称。计算所得 t 值可以是正数，也可以是负数。;

② 与 Z 分布（即标准正态分布）相比，曲线最高处较矮，两尾部翘得较高。;

③ t 分布曲线是一簇曲线，其形态变化与自由度的大小有关，自由度一旦确定，则 t 分布的形状也就确定。自由度越小，则 t 值越分散，曲线越低平；随着自由度的增大，t 分布曲线逐渐接近 Z 分布曲线，t 分布的极限分布为标准正态分布（即 Z 分布）。

④ t 分布曲线下面积有一定的规律性。



自由度$\nu=9$ 时，$t\le -1.833$ 或 $t\ge 1.833$ 的（单侧）曲线下面积为0.05；

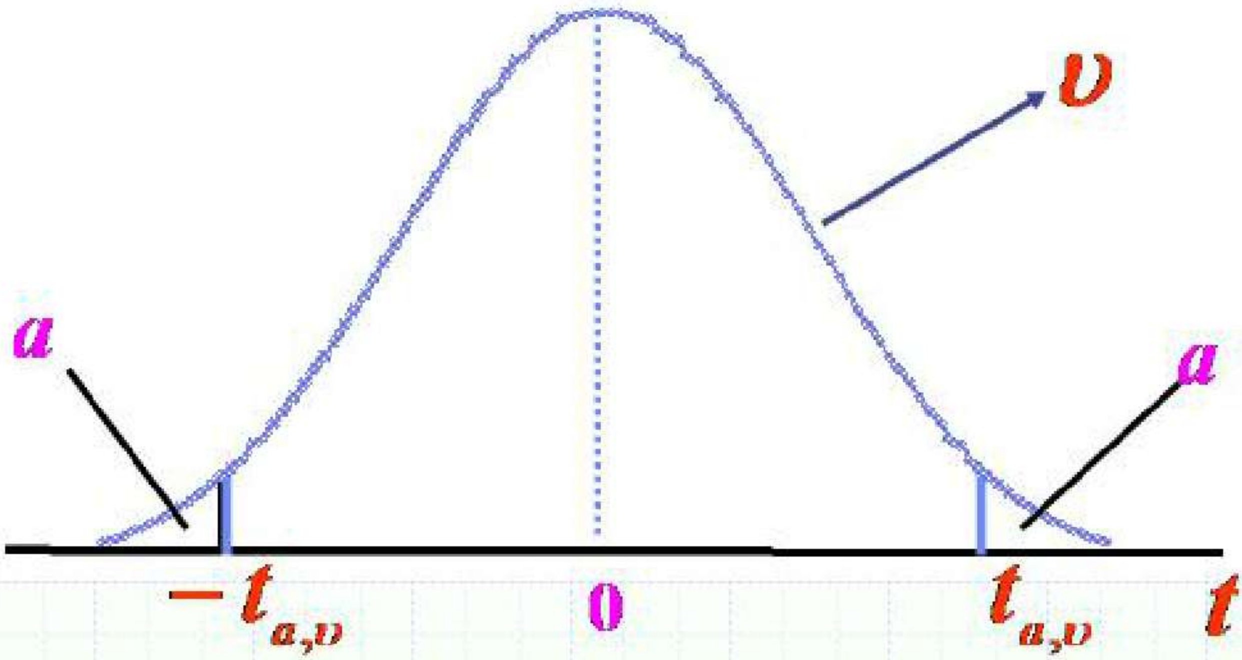
$t\le -2.262$ 和 $t\ge 2.262$ 的（双侧）曲线下面积也为0.05

三、t 界值表 p_{275}

详见附表2，可反映t分布曲线下的面积。

单侧概率或单尾概率：用$t_{\alpha,\nu}$表示；

双侧概率或双尾概率：用$t_{\alpha/2,\nu}$表示。

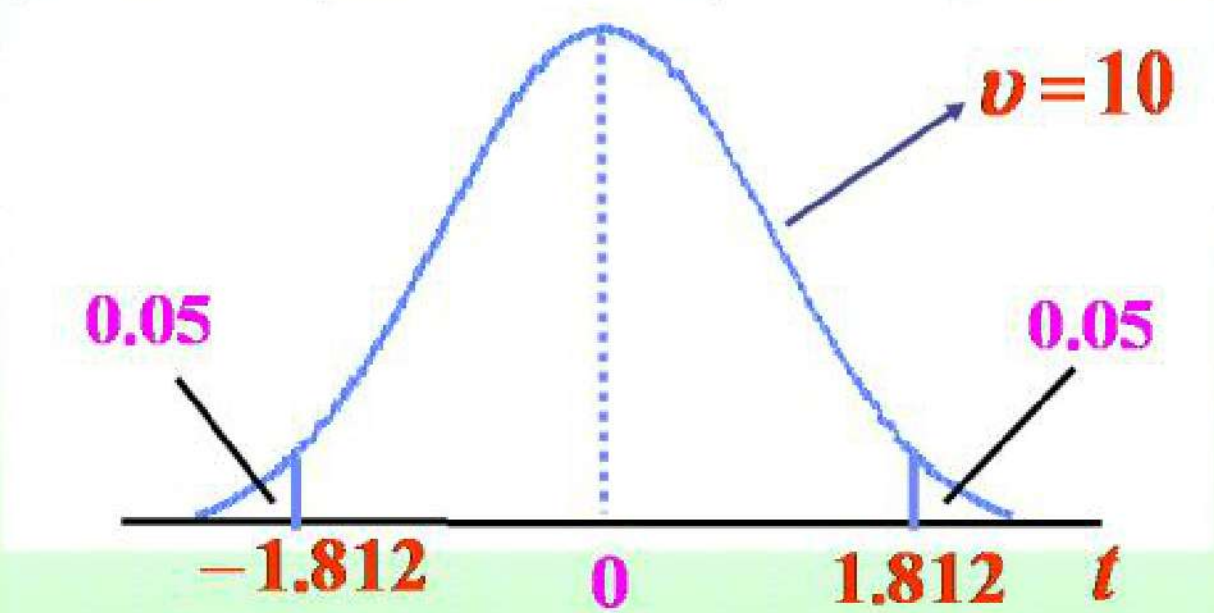


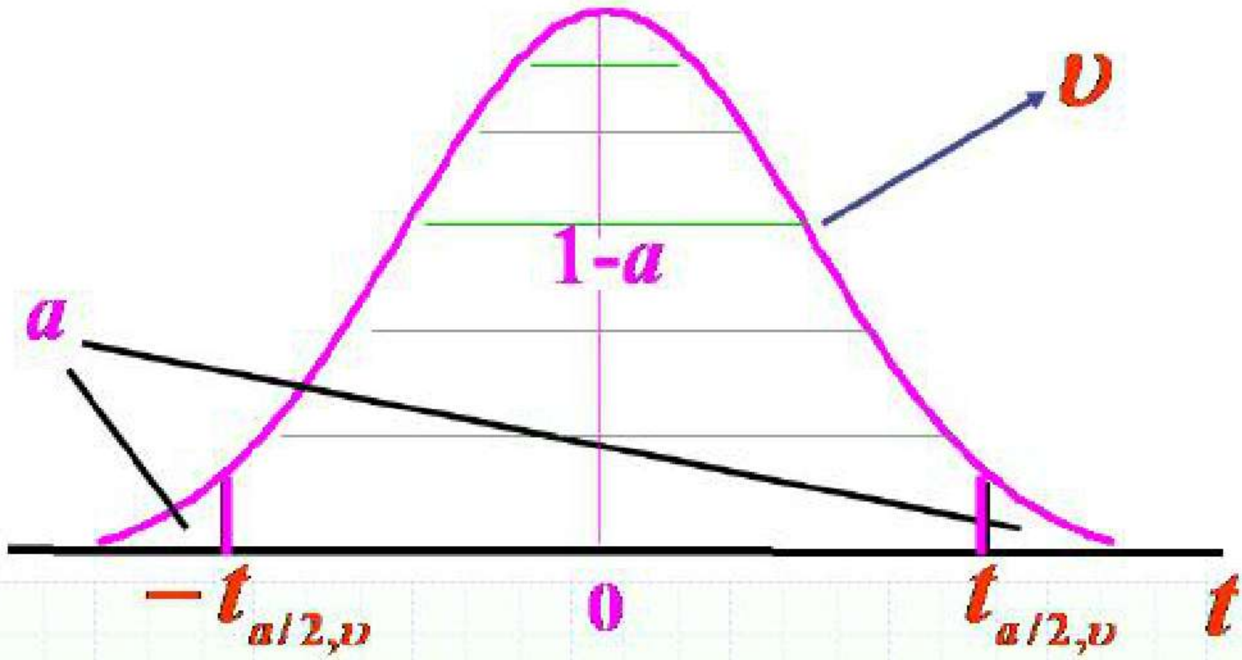
单侧概率或单尾概率：用 $t_{\alpha,\upsilon}$ 表示。

举 例：

$\nu=10$，单α=0. 05，$t_{0.05,10}=1.812$ 则有：

$P(t\leq-1.812)=0.05$ 或 $P(t\geq1.812)=0.05$



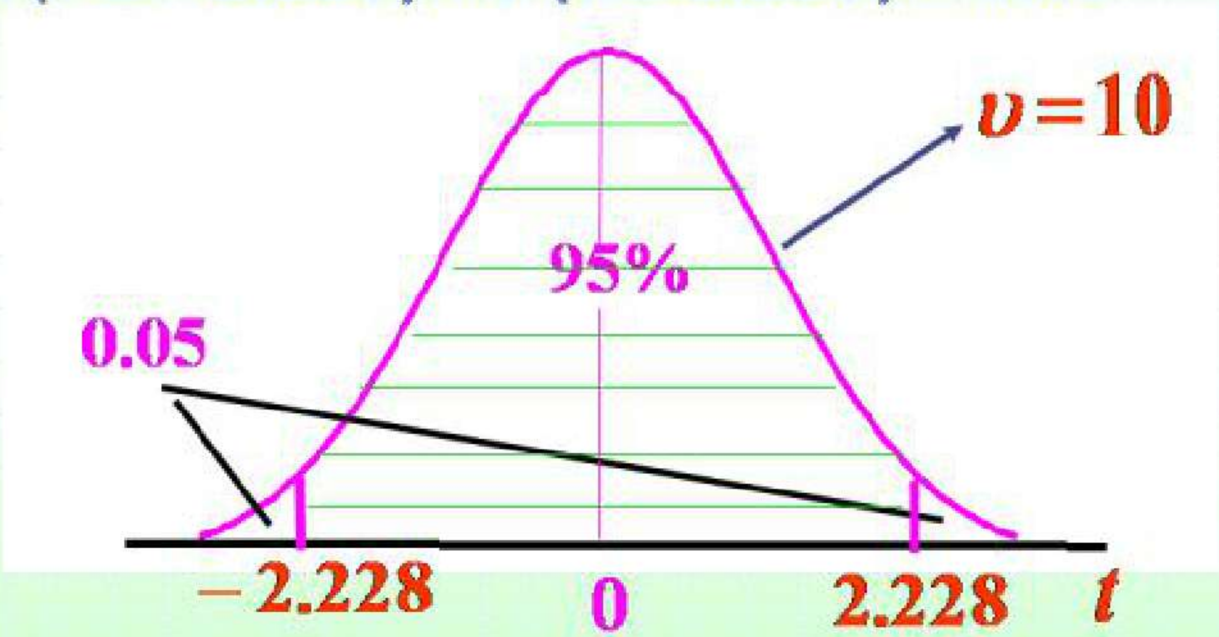


双侧概率或双尾概率：用$t_{a/2,\upsilon}$表示。

举 例：

$\nu=10$, 双α=0.05，$t_{0.05/2,10}=2.228$ 则有：

$$P(t\le -2.228)+P(t\ge 2.228)=0.05$$



t 界值表的两种用途：

1、已知自由度和概率p(单侧或双侧)，查表得单侧或双侧t界值 $t_{\alpha,\nu}$或$t_{\alpha/2,\nu}$。

2、已知自由度和统计量t值，查表得相应的概率p的大致范围。

$|t|$越大，其尾部面积即概率p越小。

第三节

总体参数的估计

一、置信区间的概念

参数估计：用样本统计量推断总体参数的大小。

包括：点（值）估计、区间估计

1. 点估计(point estimation):

就是用相应样本统计量直接作为其总体参数的估计值。如用 $\bar{X}$ 估计 μ、S 估计 σ 等。其方法虽简单，但未考虑抽样误差的大小。

2、区间估计(interval estimation):

区间估计：按预先给定的概率$(1-\alpha)$所确定的包含未知总体参数的一个范围。

- 如给定α=0.05,该范围称为总体参数的95%可信区间或置信区间；
- 如给定α=0.01,该范围称为总体参数的99%可信区间或置信区间。

二、总体均数的可信区间

1、概念：按预先给定的概率$(1-\alpha)$所确定的包含未知总体均数的一个可能范围。

2、计算：有两种方法

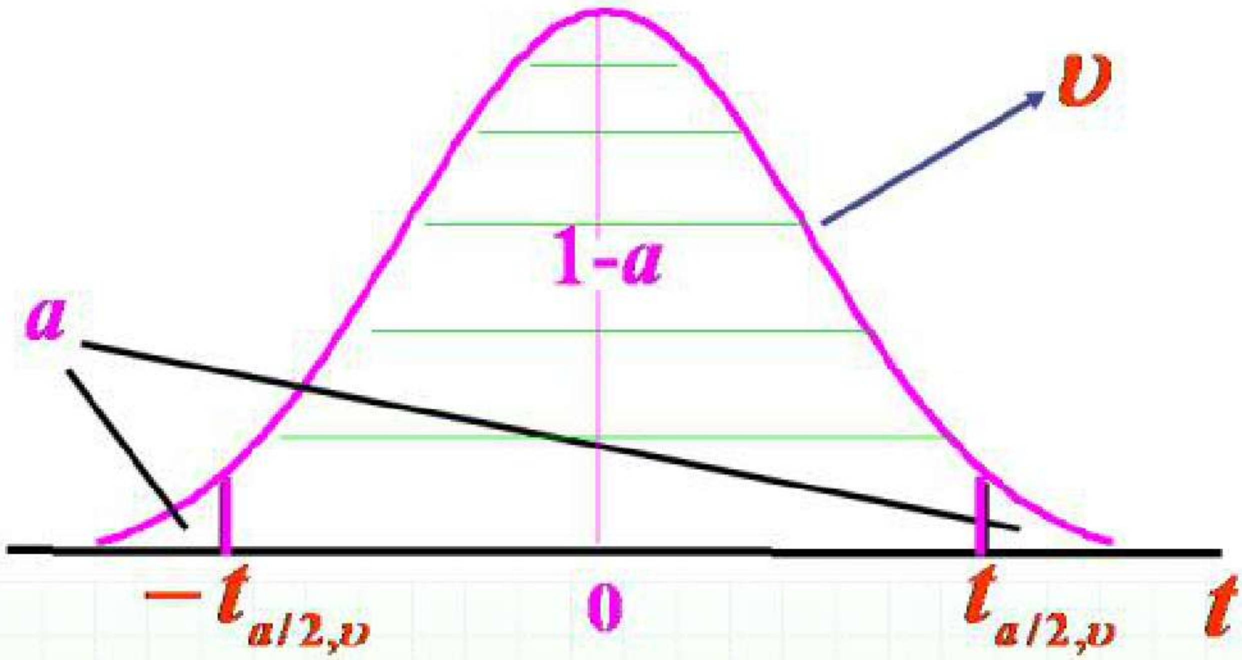
(1) σ未知时，按 t 分布原理

$$p\left(-t_{a/2,\upsilon} \prec t\left(=\frac{\overline{X}-\mu}{S_{\overline{X}}}\right) \prec t_{a/2,\upsilon}\right)=1-a$$

即 $-t_{a/2,\upsilon} \prec \dfrac{\overline{X}-\mu}{S_{\overline{X}}} \prec t_{a/2,\upsilon}$ 成立的可能性为 $1-a$

即 $\overline{X}-t_{a/2,\upsilon}S_{\overline{X}} \prec \mu \prec \overline{X}+t_{a/2,\upsilon}S_{\overline{X}}$

即 $(\overline{X}-t_{a/2,\upsilon}S_{\overline{X}}, \overline{X}+t_{a/2,\upsilon}S_{\overline{X}})$



双侧概率或双尾概率：用 $t_{\alpha/2,\upsilon}$ 表示。

总体均数双侧1-a 置信区间可简写为:

$$\bar{X} \pm t_{a/2,\upsilon} S_{\bar{X}}$$

总体均数μ的双侧95%可信区间为:

$$\bar{X} \pm t_{0.05/2,\upsilon} S_{\bar{X}}$$

总体均数单侧1-a 置信区间可简写为:

$$>\bar{X}-t_{a,\nu}S_{\bar{X}} \qquad 或<\bar{X}+t_{a,\nu}S_{\bar{X}}$$

例 7-4　在例 7-1 中抽得第 2 号样本的均数 $\overline{X}=5.03$ (cm)，标准差 S=0.52(cm)，求其总体均数的 95%可信区间。

本例 $n=10$， $\nu=n-1=10-1=9$,双尾 $\alpha=0.05$，

查附表 2 的 t 界值表得 $t_{0.05/2,9}=2.262$ 。

按公式 $5.03-2.262\times\frac{0.52}{\sqrt{10}}<\mu<5.03+2.262\times\frac{0.52}{\sqrt{10}}$

即 $(4.66, 5.40)10^{12}/L$

故该地成年男子红细胞数均数的 95%可信区间为 $(4.66, 5.40)10^{12}/L$。

可信区间的确切涵义

当$1-\alpha=95\%$时，在算得的100个可信区间中，有95个可信区间包含了总体均数，而另外5个(模拟抽样中的第3号、4号、54号、76号和82号)不包括总体均数在内。

可信区间的确切涵义:

如果能够进行重复抽样试验，平均有$1-\alpha$(如95%)的可信区间包含了总体参数。但在实际工作中，只能根据一次试验结果估计可信区间，如上例，95%的可信区间为4.66~ 5.40(10^{12}/L)，就认为该区间包含了总体均数μ (实际上也包含了总体均数在内)。

(2) n>60时，按 z 分布原理

随着样本含量 n（即自由度υ）的增大，t分布$\rightarrow z$分布，$t_{\alpha/2,\upsilon} \rightarrow z_{\alpha/2}$，可得$z$分布原理的可信区间：

$$(\overline{X} - z_{\alpha/2}S_{\overline{X}}, \overline{X} + z_{\alpha/2}S_{\overline{X}})$$

总体均数双侧1-a 置信区间可简写为:

$$\bar{X} \pm z_{a/2} S_{\bar{X}}$$

总体均数μ的95%可信区间为:

$$\bar{X} \pm 1.96 S_{\bar{X}}$$

总体均数单侧**1-α** 置信区间可简写为:

$$>\bar{X}-z_{\alpha}S_{\bar{X}} \qquad \text{或} < \bar{X}+z_{\alpha}S_{\bar{X}}$$

例　某地抽取正常成年人200名，测得其血清胆固醇的均数为3.64 mmol/L，标准差为1.20mmol/L，估计该地正常成年人血清胆固醇均数的95%可信区间。

本例 n=200>60，故可采用正态近似的方法按公式计算可信区间。今 $\bar{X}$=3.64、S=1.20、n=200、$S_{\bar{X}}$=0.0849，α 取双尾 0.05 得 $u_{0.05/2}=1.96$。

$$(3.64 \pm 1.96 \times 0.0849) = (3.47, 3.81)(\text{mmol/L})$$

故该地正常成年人血清胆固醇均数的双侧95%可信区间为(3.47, 3.81)mmol/L。

可信区间估计的优劣取决于两个方面:

- 一是可信度$1-\alpha$，愈接近1愈好，即99%的可信度比95%的可信度要好;
- 二是区间的宽度，区间愈窄愈好，即95%的可信区间比99%的可信区间要好;

在可信度确定的情况下，增加样本含量可减小区间宽度。

三、两总体均数差值的置信区间

将$(\overline{X}_1-\overline{X}_2)$看成$\overline{X}$，将$(\mu_1-\mu_2)$看成$\mu$，参照前面总体均数的置 信区间估计方法。

方法：小样本，t分布（$n_1<30$或$n_2<30$，两样本来自正态总体且两总方差齐性）

大样本，z分布（$n_1\geq30$,和$n_2\geq30$）

1、小样本，用t分布原理

双侧1-α可信区间 $(\bar{X}_1-\bar{X}_2)\pm t_{\alpha/2,\nu}S_{\bar{X}_1-\bar{X}_2}$

$$S_{\bar{X}_1-\bar{X}_2}=\sqrt{S_C^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$S_C^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$

例4-2 为了解氨甲喋呤对外周血IL-2水平的影响，某医生将60名哮喘患者随机分为两组。其中对照组29例(n_1)，采用安慰剂；实验组31例(n_2)，采用小剂量氨甲喋呤进行治疗。测得对照组治疗前IL-2的均数为20.00 IU/ml ($\bar{X}_1$)，标准差为7.00 IU/ml (S_1)；试验组治疗前IL-2的均数为17.00 IU/ml ($\bar{X}_2$)，标准差为8.50 IU/ml (S_2)。问两组总体均数相差有多大？

解：先计算 $S_{\bar{X}_1-\bar{X}_2}$，再计算95%可信区间

$$S_{\bar{X}_1-\bar{X}_2}=\sqrt{\frac{(29-1)\times 7.00^2+(31-1)\times 8.50^2}{29+31-2}\left(\frac{1}{29}+\frac{1}{31}\right)}=2.0181$$

$$\alpha=0.05 \qquad \nu=n_1+n_2-2=29+31-2=58$$

查附表二，t界值表得 $t_{0.05/2,60}=2.000$

$$(20.0-17.0)\pm 2.000\times 2.0181=(-1.0362,7.0362)$$

即两组治疗前的IL-2总体均数之差的95%置信区间为(-1.04，7.04) (IU/mL) 。

2、大样本，用z分布原理

双侧1-α可信区间：$(\bar{X}_1-\bar{X}_2)\pm z_{\alpha/2}S_{\bar{X}_1-\bar{X}_2}$

$$S_{\bar{X}_1-\bar{X}_2}=\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

双侧95%可信区间：$(\bar{X}_1-\bar{X}_2)\pm 1.96S_{\bar{X}_1-\bar{X}_2}$

双侧99%可信区间：$(\bar{X}_1-\bar{X}_2)\pm 2.58S_{\bar{X}_1-\bar{X}_2}$

四、总体率的置信区间

与总体均数的估计基本相似，包括总体率的点估计和区间估计，而区间估计又有两种方法。

方法：小样本，查表法（$n \le 50$，二项分布）

大样本，z分布（$np \ge 5$且$n(1-p) \ge 5$）

$$(p - Z_{\alpha/2}S_p, p + Z_{\alpha/2}S_p)$$

例4-3

2003年4～6月某医院重症监护病房收治重症SARS患者38人，其中死亡14人，求SARS病死率的置信区间。

解：

查附表3，在$n=38\ X=14$ 的纵横交叉处的数值是上行22－54，下行18－59，即SARS病死率的95%的置信区间为22%～54%，99%的置信区间为18%～59%。

- 注意：附表3中的X只列出部分，当时，应以值查表，然后用100减去查表得的数值，即为所求的置信区间。

例4-4

甘肃省天水市北道区卫生防疫站2002年对该乡镇250名小学生进行贫血的检测，结果发现有86名贫血者，检出率为34.4%，求贫血检出率的95%的置信区间。

解:

本例n比较大，且 $np=86$ 及 $n(1-p)=164$ 均大于5，所以可用公式估计总体率的置信区间。

$$p \pm Z_{\alpha/2}S_p = p \pm Z_{0.05/2}\sqrt{\frac{p(1-p)}{n}} = 0.344 \pm 1.96\sqrt{\frac{0.344(1-0.344)}{250}} = (0.285, 0.403)$$

即该乡镇小学生贫血的检出率的**95**%置信区间为**(23.5**%，**34.7**%**)**。

五、两总体率差值的置信区间

与两总体均数差值的置信区间估计基本相似。

条件是：（$n_1 p_1 \geq 5$、$n_1(1-p_1) \geq 5$

和（$n_2 p_2 \geq 5$、$n_2(1-p_2) \geq 5$

方法：$(p_1 - p_2) \pm z_{\alpha/2} S_{p_1 - p_2}$

例4-5 某医院口腔科医生用极固宁治疗牙本质过敏症，以双氟涂料作对照，进行了1年的追踪观察，结果如表4-4。试估计两组有效率差别的95%置信区间。

表4-4 两组效果比较

组别	总牙数	有效数	有效率（%）
实验组	77	61	79.2
对照组	69	38	55.1

解：$n_1=77, X_1=61, p_1=79.2; n_2=69, X_2=38, p_2=55.1$

1. 计算两样本的合并有效率，得

$$p_c=\frac{X_1+X_2}{n_1+n_2}=\frac{61+38}{77+69}=67.8\%$$

2. 用公式(4-14)计算两率之差的标准误

$$S_{p_1-p_2}=\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}=\sqrt{0.687(1-0.687)\left(\frac{1}{77}+\frac{1}{69}\right)}=0.077=7.7\%$$

3. 用公式(4-22)计算两组总体率有效率差别的95％置信区间

$$(p_1-p_2)\pm z_{\alpha/2}S_{p_1-p_2}=(79.2-55.1)\pm1.96\times7.7=(9.0\%,39.2\%)$$

4. 推断 该区间下限值大于0，即没有包括，所以两组治疗效果的总体有效率不同，结合样本率可知，极固宁治疗组的总体有效率高于对照组