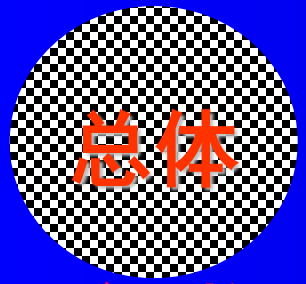
参数估计基础

抽样研究的目的就是要用样本信息来推断相应总体的特征，这一过程称为统计推断。

统计推断包括两方面的内容：参数估计和假设检验

统计推断
statistical inference



抽取部分观察单位

参 数

统计推断

统计量

如：总体均数 μ

总体标准差 σ

总体率 π

如：样本均数 $\bar{X}$

样本标准差S

样本率 P

内容：

1. 参数估计 (estimation of parameters)

   包括：点估计与区间估计

2. 假设检验（test of hypothesis)

误差：泛指测得值与真值之差，样本指标与总体指标之差。误差按其产生的原因与性质分为两大类（系统误差和偶然误差）。

1.系统误差：由于受试对象、研究者、仪器设备、研究方法、非实验因素影响等确定性原因造成，有一定倾向性或规律性的误差。可以避免。

2.随机测量误差：由于多种无法控制的偶然因素引起，对同一样品多次测量数据的不一致。无倾向性，不可避免。只可控制在一定的范围内。

3.抽样误差：由个体变异产生的、由于抽样而造成的样本统计量与样本统计量及样本统计量与总体参数之间的差异称为抽样误差。无倾向性，不可避免。

均数的抽样误差、总体均数的估计、t分布

1、均数的抽样误差和标准误

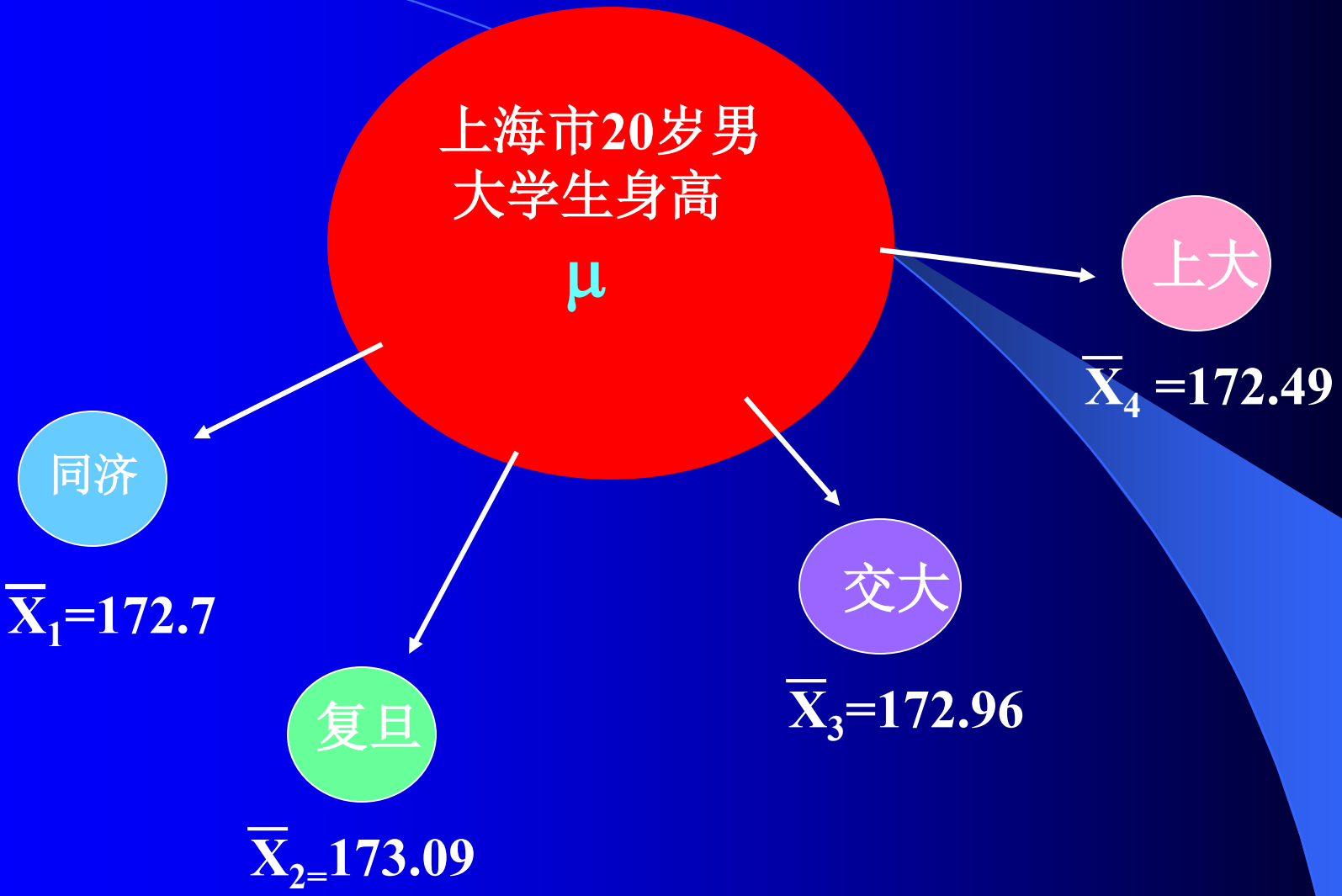
抽样试验

以110名20岁健康男大学生的身高作为假设的有限总体，其总体均数 $\mu = 172.73(cm)$ 标准差 $\sigma = 4.09(cm)$

每次随机抽取10个人的身高作为一个样本，记录下数据并计算均数、标准差，再放回重新抽样，共重复100次，求得100个样本均数和标准差，其样本均数列入表3.1。

表3.1 100个样本均数

173.22	172.06	170.89	174.07	172.60	173.14	172.61	172.26	171.93	172.85
175.23	173.76	174.77	172.57	171.76	172.74	173.36	173.69	171.10	173.40
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
172.70	172.77	173.47	172.13	172.56	172.13	169.63	170.71	172.63	172.14


上海市20岁男
大学生身高
μ
上大
$\overline{X}_4$ =172.49
同济
$\overline{X}_1$=172.7
交大
$\overline{X}_3$=172.96
复旦
$\overline{X}_{2=}$173.09

例3-1 某市1999年18岁男生身高服从μ=167.7cm、σ=5.3cm正态分布，从该$N(167.7, 5.3^2)$总体中随机抽样。

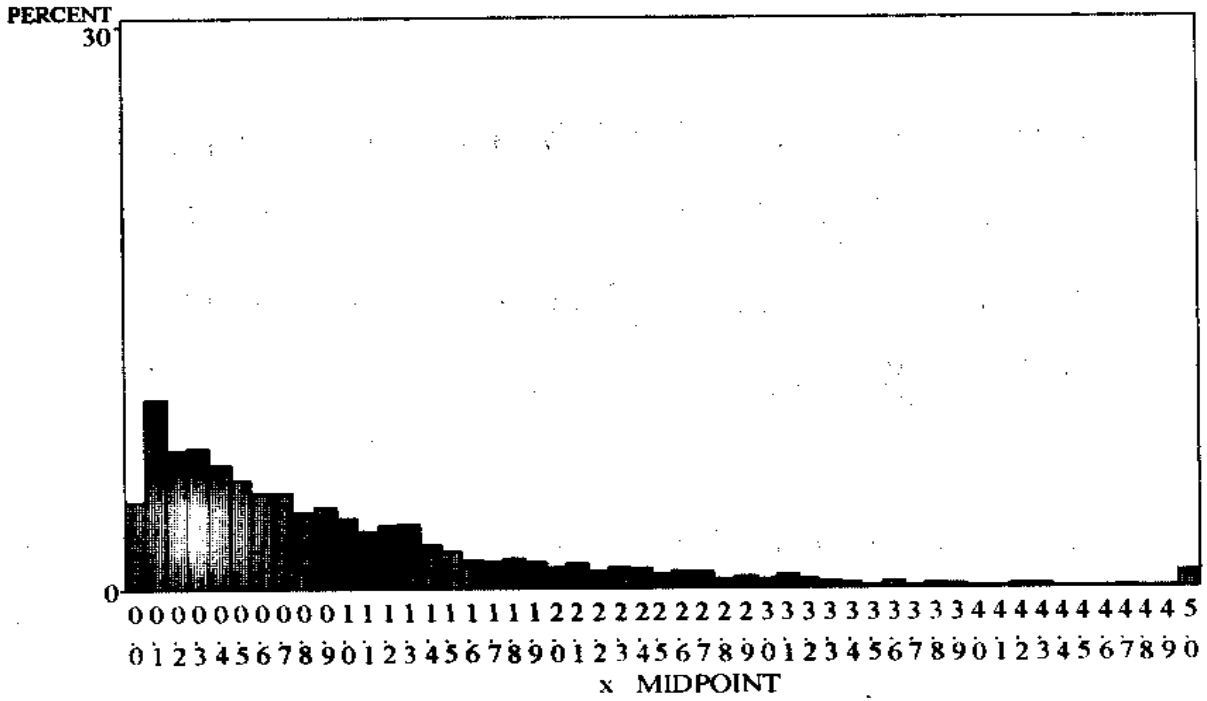
每次n_j=10人，共有样本g=100个，得到每个样本均数$\overline{X}_j$及标准差S_j。

将上述100个样本均数看成新变量值，这100个样本均数构成一新分布。

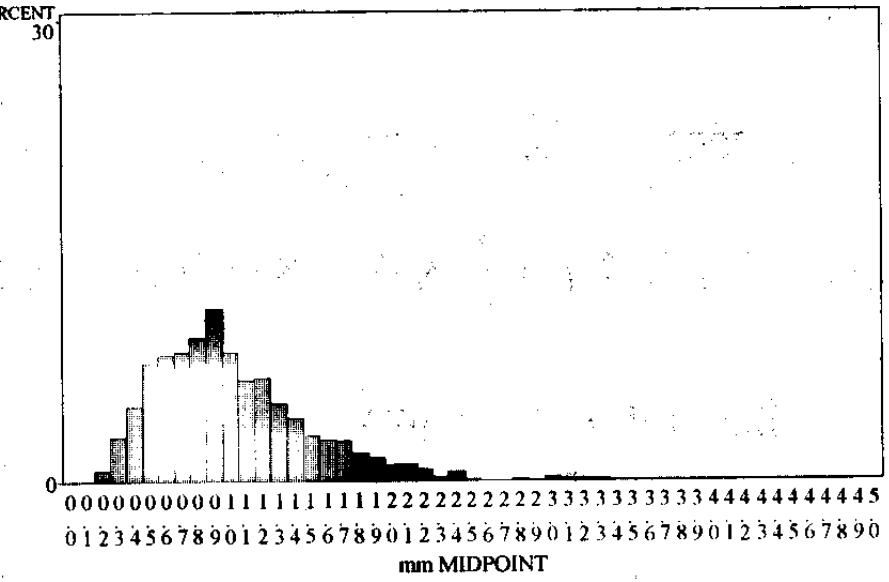
样本均数抽样分布具有如下特点：

①各样本均数未必等于总体均数；
②各样本均数间存在差异；
③样本均数围绕总体均数(167.7cm)
呈正态分布；
④样本均数变异范围较原变量变异范
围大大缩小，这100个样本均数的
均数为167.69cm、标准差为1.69cm。

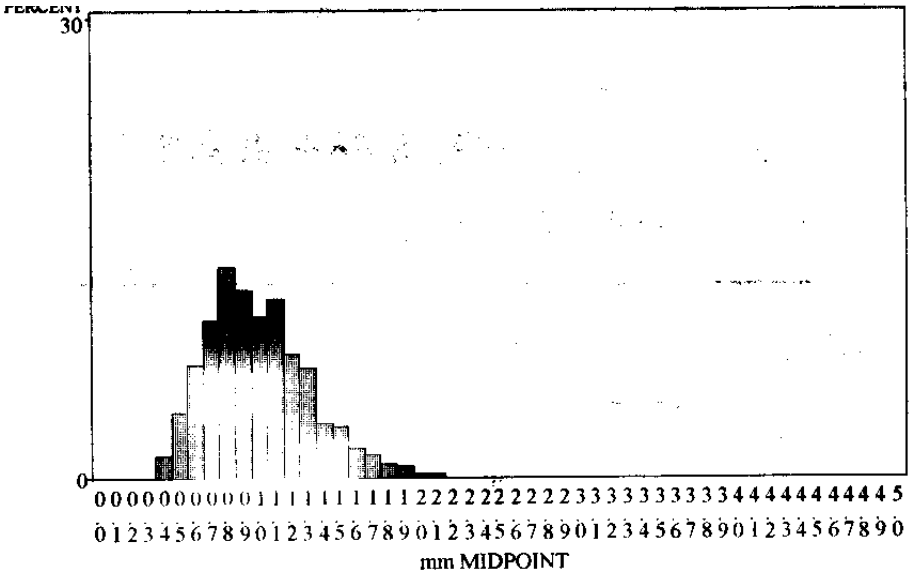
在非正态分布总体中可进行类似抽样。



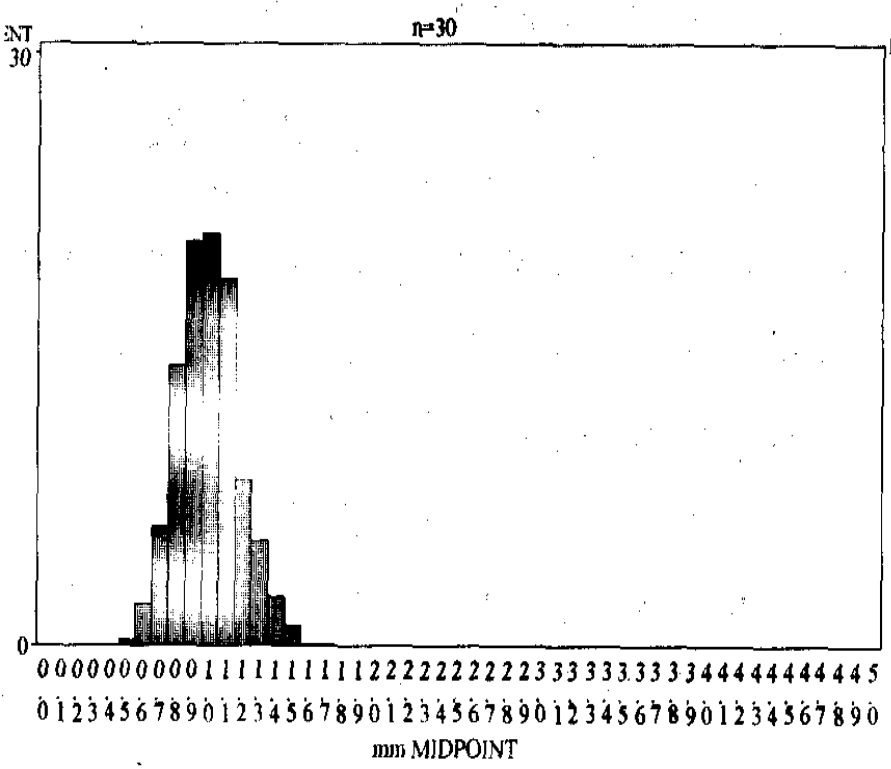
(a) 原始数据



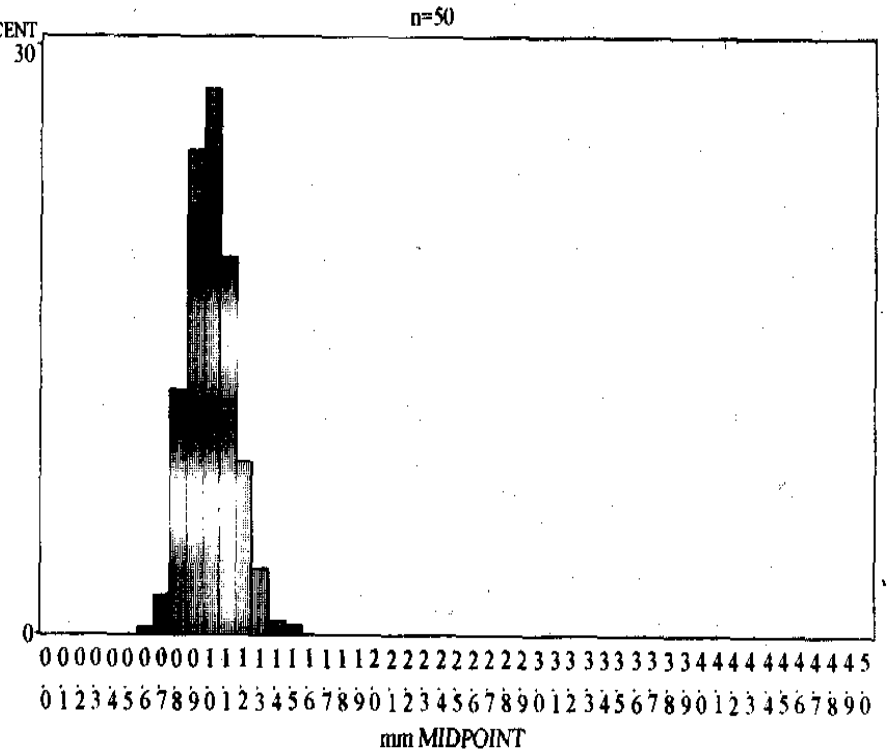
(b) n=5



(c) n=10



(d) n=30



(e) n=50

图 5-1　从正偏峰总体抽样,样本均数的分布

(a)为正偏峰总体,(b)～(e)为不同样本含量时样本均数的直方图

数理统计推理和中心极限定理表明:

- 从$N(\mu,\sigma^2)$中随机抽取n例的样本，样本均数 $\bar{X}$也服从正态分布,且 $\bar{x} \sim N(\mu,\sigma_{\bar{x}}^2)$
- 即使从非正态总体中抽取样本，当n足够大（n>30），$\bar{X}$分布仍近似正态分布。
- 随着样本量的增大，样本均数的变异范围也逐渐变窄。

- 均数的抽样误差(sampling error of mean)

----由于抽样而造成的样本均数与总体均数之差异或各样本均数之差异称为均数的抽样误差。

- 标准误(standard error)

----反映均数抽样误差大小的指标是样本均数 的标准差，简称标准误。

标准误的计算

- 理论值 $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

- 估计值 $s_{\bar{x}} = \dfrac{s}{\sqrt{n}}$

标准误用于说明抽样误差的大小。

实例分析

例如：某地成年男子红细胞数的抽样调查，

$n=114$人，$\overline{X}=5.38\times10^{12}/L$，$s=0.44\times10^{12}/L$,

求其标准误。

$$s_{\bar{x}}=\frac{s}{\sqrt{n}}=\frac{0.44}{\sqrt{144}}=0.037(10^{12}/L)$$

例： 2000年某研究所随机调查某地健康成年男子27人，得到血红蛋白的均数为125g/L，标准差为15g/L 。试估计该样本均数的抽样误差。

$$S_{\bar{X}} = S / \sqrt{n} = 15 / \sqrt{27} = 2.89 g / l$$

2　样本频率的抽样分布与抽样误差

从同一总体中随机抽出观察单位相等的多个样本，样本率与总体率及各样本率之间都存在差异，这种差异是由于抽样引起的，称为**频率的抽样误差**。

表示频率的抽样误差大小的指标叫**频率的标准误**。

据数理统计的原理，率的标准误用σ_P表示

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

：总体率，n：样本例数。

当π未知时，p —π（为样本含量足够大，且p和1-p不太小）

公式为:

$$S_P = \sqrt{\frac{P(1-P)}{n}}$$

S_P：率的标准误的估计值，p：样本率。

例：　某市随机调查了50岁以上的中老年妇女776人，其中患有骨质疏松症者322人，患病率为41.5%，试计算该样本频率的抽样误差。

$$S_P = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.415(1-0.415)}{776}} = 0.0177 = 1.77\%$$

标准差和标准误的比较

	标准差	标准误
1.概念不同	是衡量个体观察值变异程度的指标。	描述样本均数距总体均数的离散程度，是抽样误差大小的尺度。
2.用途不同	A.衡量均数的代表性 B.与均数结合估计正常值范围 C.计算变异系数和标准误	A.衡量样本均数代表总体均数的可靠性 B.与样本均数结合估计总体均数的可信区间 C.进行显著性检验
3.与例数的关系不同	当样本含量足够大时，标准差趋于稳定。	随例数的增大而减小，若样本含量趋于总体例数，则标准误近似零，即抽样误差为零。
4.相同之处	变异指标 （个体观察值距样本均数）	变异指标 （样本均数距总体均数）

2、t 分布（distribution）

t分布最早
由英国统计学家W.S. Gosset
于1908年
以“Student”笔名发表，
故又称Student’s t-distribution。

它的发现，
开创了小样本统计推断的新纪元。

随机变量X

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

$$u=\frac{X-\mu}{\sigma}$$

u 变换

标准正态分布

$N(0,1)$

均数 $\overline{X}$

$N(\mu,\sigma_{\overline{x}}^2)$

$$u=\frac{\overline{X}-\mu}{\sigma_{\overline{x}}}$$

标准正态分布

$N(0,1)$

t变换

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

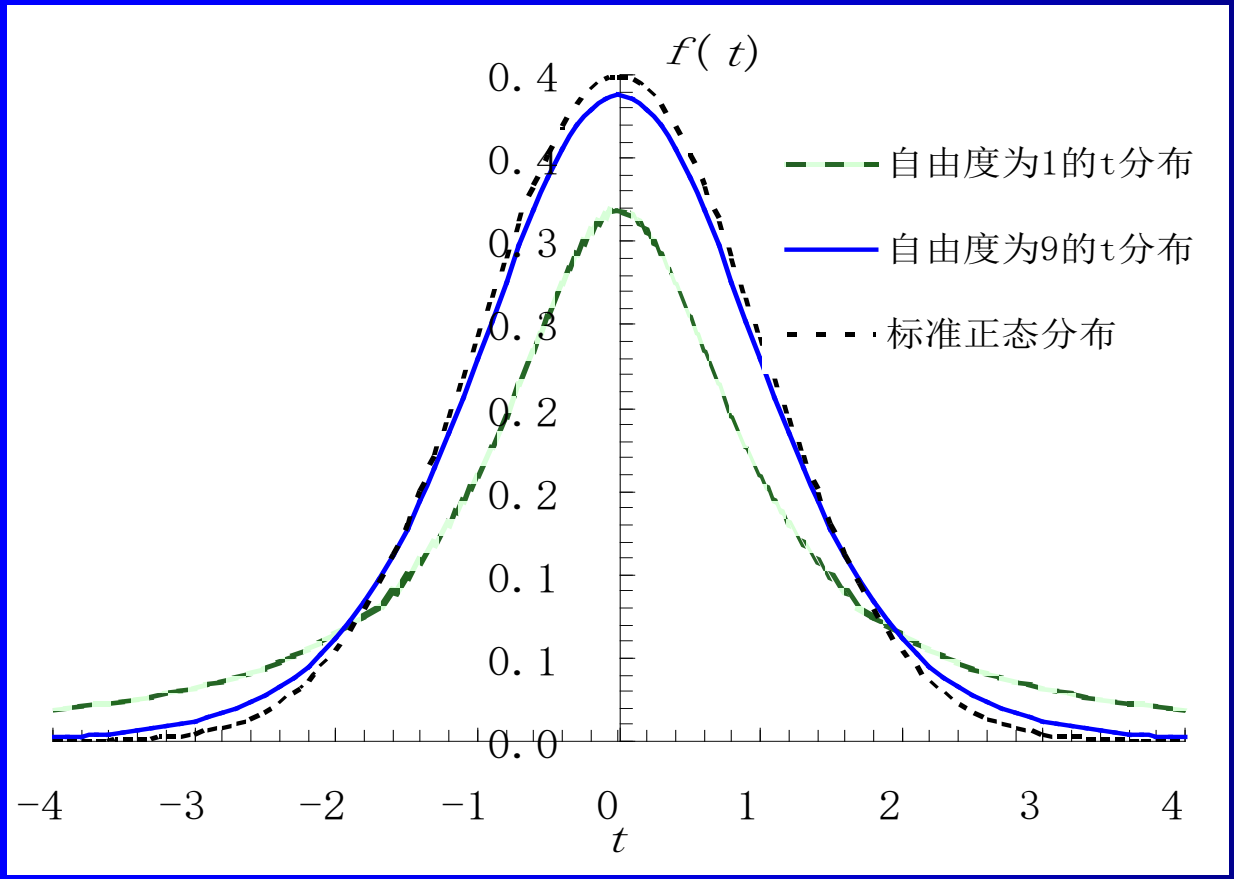
Student t分布

自由度：$n-1$

t 分布曲线

t 分布的特征：

①以t=0为中心的对称分布，曲线在t=0 处最高



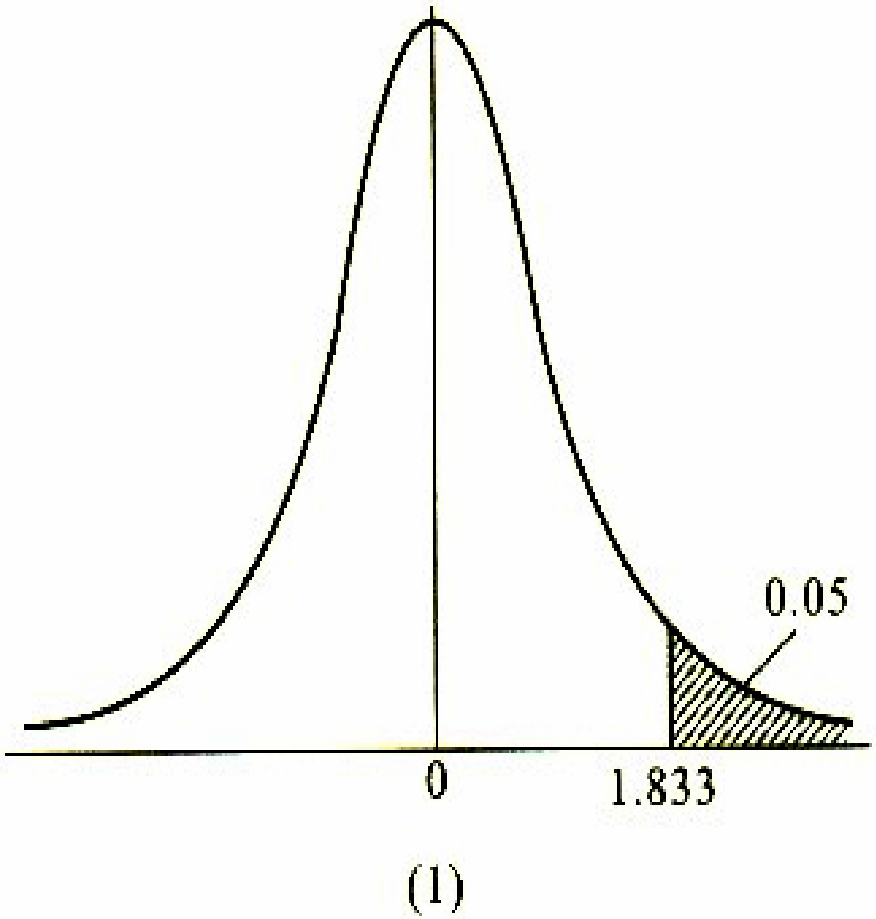
②与正态分布相比，曲线最高处较矮，两尾部翘得高（见绿线）

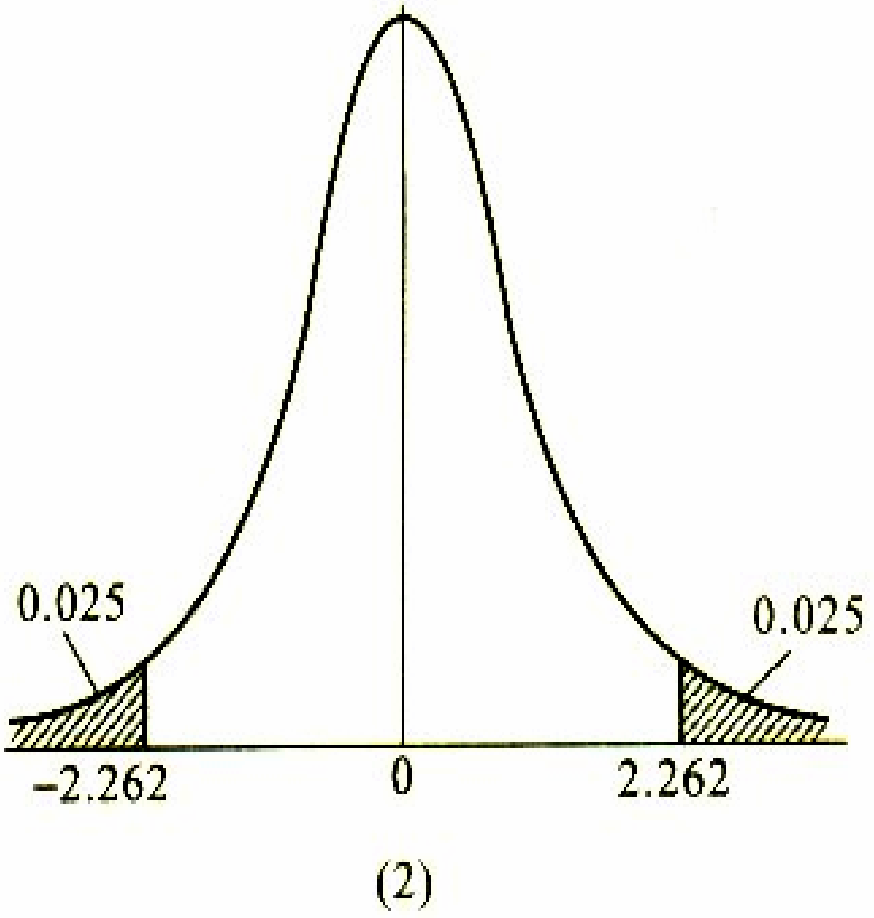
③ 随自由度增大，曲线逐渐接近正态分布；分布的极限为标准正态分布。

t 界值表

- 根据不同自由度时曲线下面积与t值的关系编制
- 横标目为自由度 υ，纵标目为概率P，表中数字为 $t_{\alpha,\upsilon}$。
- 表中只列出正值，负的*t* 值用其绝对值查表。
- 相同自由度时, *t* 的绝对值越大,概率P越小.
- 相同 υ、相同*t* 界值下，双侧P为单侧P的两倍。

$\nu=9$





t分布与 u分布的异同

- 相同点：1.二者都是以0为中心，左右对称。
  2.已平均数为最高点向两侧逐渐降。
  3.尾部无限延伸，不与基线相交。
- 不同点：1. t分布峰部矮尖尾翘，尤其自由度小时更为明显。当自由度逐渐增时，t 分布逐渐逼近u分布。
  2.标准正态分布（u分布）曲线下面积95%和99%的界值是一个常量，而t分布曲线不是常量，而是随着自由度大小而变化。

3、总体均数可信区间的估计

参数估计是指用样本指标(统计量)估计总体指标(参数).

参数估计

点值估计（point estimation）

将样本指标作为总体指标的估计值。

区间估计（interval estimation ）

按一定的可信度估计总体均数所在的范围。

1.点估计

用样本统计量直接作为总体参数的估计值。

例如 于2000年测得某地27例健康成年男性血红蛋白量的样本均数为125g/L，试估计其总体均数。

$\bar{X} \to \mu$，即认为2000年该地所有健康成年男性血红蛋白量的总体均数为125g/L 。

同理，例5-2中776名50岁以上的中老年妇女骨质疏松症的样本患病率作为总体患病率的点值估计值，即认为该市所有50岁以上的中老年妇女骨质疏松症的总体患病率约为41.5%。

1. 单一总体均数的可信区间

(1)σ未知：

双侧1－ α 可信区间

$$\overline{X}-t_{\alpha/2,\upsilon}S_{\overline{X}},\overline{X}+t_{\alpha/2,\upsilon}S_{\overline{X}}$$

单侧1－ α 可信区间

$$\mu>\overline{X}-t_{\alpha,\nu}S_{\overline{X}}$$

$$\mu<\overline{X}+t_{\alpha,\nu}S_{\overline{X}}$$

例3-2　在例3-1中抽得第15号样本的 $\overline{X}$ =166.95(cm)， S=3.64(cm)，求其总体均数的95%可信区间。

$$S_{\bar{X}} = \frac{3.64}{\sqrt{10}} = 1.1511 \ \textbf{(cm)} \qquad t_{0.05/2,9} = 2.262$$

$$(166.95 - 2.262 \times 1.1511,\ 166.95 + 2.262 \times 1.1511)$$

该故地18岁男生身高均数的95%可信区间为(164.35, 169.55)cm。

- 例：　已知某地27例健康成年男性血红蛋白量的均数为 $\overline{X}=125g/L$，标准差S=15g/L，试问该地健康成年男性血红蛋白量的95%和99%置信区间。

本例n=27，S=15

$$\bar{X} \pm t_{0.05/2(26)} \frac{15}{\sqrt{27}} = 125 \pm 2.056 \times 2.38$$

$$(119.06, 130.94)$$

$$\bar{X} \pm t_{0.01/2(26)} \frac{15}{\sqrt{27}} = 125 \pm 2.779 \times 2.38$$

$$(116.98, 133.02)$$

(2) σ已知或σ未知但n足够大:

σ已知:

双侧1– α 可信区间

$$\overline{X}-u_{\alpha/2,\upsilon}\sigma_{\overline{X}},\overline{X}+u_{\alpha/2,\upsilon}\sigma_{\overline{X}}$$

单侧1– α 可信区

$$\mu>\overline{X}-u_{\alpha/2,\upsilon}\sigma_{\overline{X}}$$

$$\mu<\overline{X}+u_{\alpha/2,\upsilon}\sigma_{\overline{X}}$$

σ未知但n足够大：

双侧1– α 可信区间

$$\overline{X}-u_{\alpha/2,\upsilon}S_{\overline{X}},\overline{X}+u_{\alpha/2,\upsilon}S_{\overline{X}}$$

单侧1– α 可信区间

$$\mu>\overline{X}-u_{\alpha/2,\upsilon}S_{\overline{X}}$$

$$\mu<\overline{X}+u_{\alpha/2,\upsilon}S_{\overline{X}}$$

例3-3　某地抽取正常成年人200名，测得其血清胆固醇均数为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95%可信区间。

- 本例 $\overline{X}$ =3.64、S=1.20、n=200、
- $S_{\overline{X}}$=0.0849，

$$u_{0.05/2} = 1.96$$

$$\left(3.64 - 1.96 \times 0.0849,\ 3.64 + 1.96 \times 0.0849\right)$$

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

例5-4 某市2000年随机测量了90名19岁健康男大学生的身高，其均数为172.2cm，标准差为4.5cm,，试估计该地19岁健康男大学生的身高的95%置信区间。

$$\overline{X} \pm z_{\alpha/2} S_{\overline{X}} \qquad z_{0.05/2} = 1.96$$

$$\overline{X} \pm 1.96 S_{\overline{X}} = 172.2 \pm 1.96 \frac{4.5}{\sqrt{90}} = (171.3, 173.1)$$

市19岁健康男大学生的身高的95%置信区间(171.3,173.1)cm

总体均数可信区间的估计

可信区间	σ已知	σ未知 但n足够大	σ未知 且n小
95% $S_{\bar{x}}$	$\bar{X} \pm 1.96\sigma_{\bar{x}}$	$\bar{X} \pm 1.96S_{\bar{x}}$	$\bar{X} \pm t_{0.05(\nu)}S_{\bar{x}}$
99% $S_{\bar{x}}$	$\bar{X} \pm 2.58\sigma_{\bar{x}}$	$\bar{X} \pm 2.58S_{\bar{x}}$	$\bar{X} \pm t_{0.01(\nu)}S_{\bar{x}}$

（二）、总体概率的置信区间

总体概率的置信区间与样本含量n，阳性频率p的大小有关，可根据n和p的大小选择以下两种方法。

1. 正态近似法

当样本含量足够大，且p和1-p不太小，则样本率的分布近似正态分布。

公式为：

$$\left(P - Z_{\alpha/2}S_P, P - Z_{\alpha/2}S_P\right)$$

S_P为率的标准误的估计值，

例5-7　用某种仪器检查已确诊的乳腺癌患者94例，检出率为78.3%。估计该仪器乳腺癌总体检出率的95%置信区间。

分析：本例样本例数较大，且样本率p不太小，可用正态近似法：

$$p \pm z_{\alpha/2} S_p$$

$$\mathrm{P} \pm \mathrm{z}_{\alpha/2} \mathrm{S}_{\mathrm{P}} = \mathrm{p} \pm \mathrm{z}_{0.05/2} \sqrt{\frac{\mathrm{p}(1-\mathrm{p})}{\mathrm{n}}}$$

$$= 0.783 \pm 1.96 \times \sqrt{\frac{0.783(1-0.783)}{120}}$$

$$= (0.709, 0.857)$$

2. 查表法

当n较小，如$n \leq 50$，特别是p和1-p接近0或1时，应按照二项分布的原理估计总体率的可信区间。

例5-5 某医院对39名前列腺癌患者实施开放手术治疗，术后有合并症者2人，试估计该手术合并症发生概率的95%置信区间。

例5-6 某医生用某药物治疗31例脑血管梗塞患者，其中25例患者治疗有效，试求该药物治疗脑血管梗塞有效概率的95%置信区间。

☞可信区间和可信限

- 可信区间（confidence interval 简记为CI）

  可信区间是以上下可信限为界的一个范围。例如95%的可信区间为（171.97，173.49）cm。

- 可信限（ confidence limit 简记为CL）

  可信限是指上限和下限两个点值。如171.97为下限

结果报告：可将点值估计和区间估计同时写出

如 172.72（171.97，173.49）cm

☞可信区间的两个要素

1、准确度——即区间包含总体均数可能性（概率）大小，反映在可信度 $1-\alpha$ 的大小。愈接近1愈好，如可信度 99%比95%好。

2、精度——反映在区间的长度，长度愈小愈精确。

在样本例数确定的情况下，二者是矛盾的。

通常要兼顾准确度与精度，因此95%CI常用。

☞医学参考值范围与可信区间的区别

❶概念（意义）不同：

95%医学参考值范围：一般是指同质总体内包括95%个体观察值的估计范围。

95%CI：是指按95%可信度估计的总体均数的所在范围。

❷公式不同：

95%医学参考值范围　$\bar{X} \pm 1.96S$

95%CI　$\bar{X} \pm 1.96S_{\bar{x}}$

❸结果不同：

95%医学参考值范围　164.71~180.70 cm

95%CI　171.97~173.49 cm