独立样本检验

		方差方程的 Levene 检验		均值方程的 t 检验						
									差分的 95% 置信区间	
		F	Sig.	t	df	Sig.(双侧)	均值差值	标准误差值	下限	上限
DNA含量	假设方差相等	.721	.412	2.395	12	.034	1.9000	.7933	.1715	3.6285
	假设方差不相等			2.395	11.566	.035	1.9000	.7933	.1642	3.6358

抽样误差

抽样误差（sampling error）：

由于个体差异的存在，导致抽样时样本统计量与总体参数间的差别；或同一总体的相同统计量之间的差别。

属于随机误差：无倾向性，不可避免。

抽样实验

假定正常成年男子的红细胞计数服从正态分布$N(5.00, 0.50^2)$，从该总体中随机抽样，样本含量$n=10$，计算其均数与标准差；重复抽取100次，获得100份样本；计算100份样本的均数与标准差，并对100份样本的均数作直方图。单位：10ˆ12/L

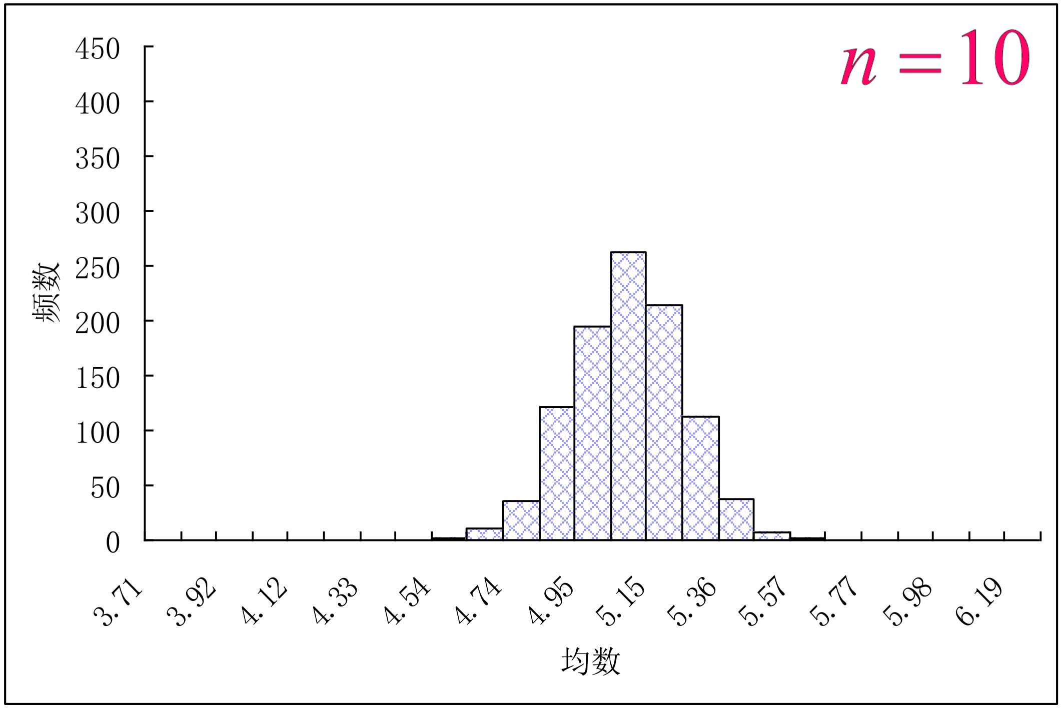
按上述方法再做样本含量$n=5$、$n=30$的抽样试验；比较计算结果。

表1　正常男子红细胞计数抽样实验结果

(μ=5.00, σ^2=0.50^2， n=10)

样本号	红细胞计数 (X) ($10^{12}/L$)					$\overline{X}$	S
1	5.59	5.11	4.26	…	5.55	5.04	0.44
2	4.65	4.65	5.59	…	5.32	5.03	0.52
3	4.56	4.87	5.21	…	4.23	4.71	0.33
…	…	…	…	…	…	…	…
100	5.16	4.49	5.26	…	4.56	4.90	0.29

样本均数和标准差



$$\overline{\overline{X}} = \frac{\sum_{i=1}^{N} \overline{X}_i}{N} = \frac{5.04 + 5.03 + 4.71 + \cdots + 4.90}{100} \approx 5.00 = \mu$$

$$S_{\overline{X}} = \sqrt{\frac{\sum (\overline{X} - \overline{\overline{X}})^2}{N-1}} = 0.1586$$

抽样实验小结

正态总体$N(\mu,\sigma^2)\rightarrow$样本均数$\overline{X}\sim N(\mu,\sigma^2/n)$

均数的均数围绕总体均数上下波动。

均数的标准差 $\sigma_{\overline{X}}$ 与总体标准差σ相差一个常数倍数，即 $\sigma_{\overline{X}}=\sigma/\sqrt{n}$ 。

样本均数的标准误

1. 标准误(*Standard Error, SE*)：统计量的标准差
2. 样本均数的标准误：样本均数的标准差，测度样本均数的抽样误差，即样本均数的离散程度。
3. 理论值：$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

   估计值：$$S_{\overline{X}} = \frac{S}{\sqrt{n}} \rightarrow = \quad SE$$
4. 小于总体标准差
5. 影响抽样误差大小的主要因素是样本量

例题

例7-3　2003年某地20岁应征男青年中随机抽取85人，平均身高为171.2cm，标准差为5.3cm，计算当地20岁应征男青年身高的标准误。

$$S_{\overline{X}} = S/\sqrt{n} = 5.3/\sqrt{85} = 0.57(cm)$$

即本次调查身高均数171.2cm抽样误差的估计值为0.57cm　。

参数估计(estimation of parameter)

参 数 估 计

点 估 计

区间估计

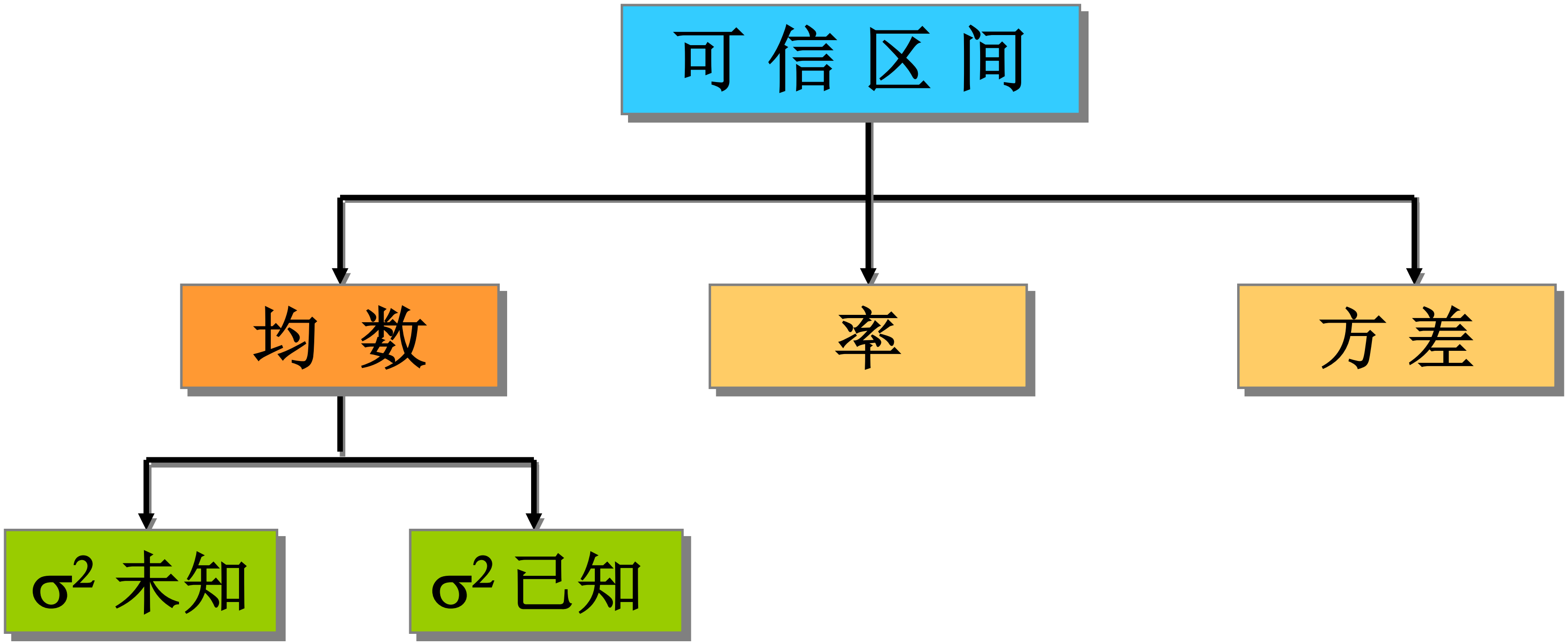
样本统计量 $\overline{X}$、S、p

总体参数 μ、σ、π

利用样本信息计算一个区间，并给出重复试验时该区间包含总体参数的概率

可信区间(confidence interval, CI)

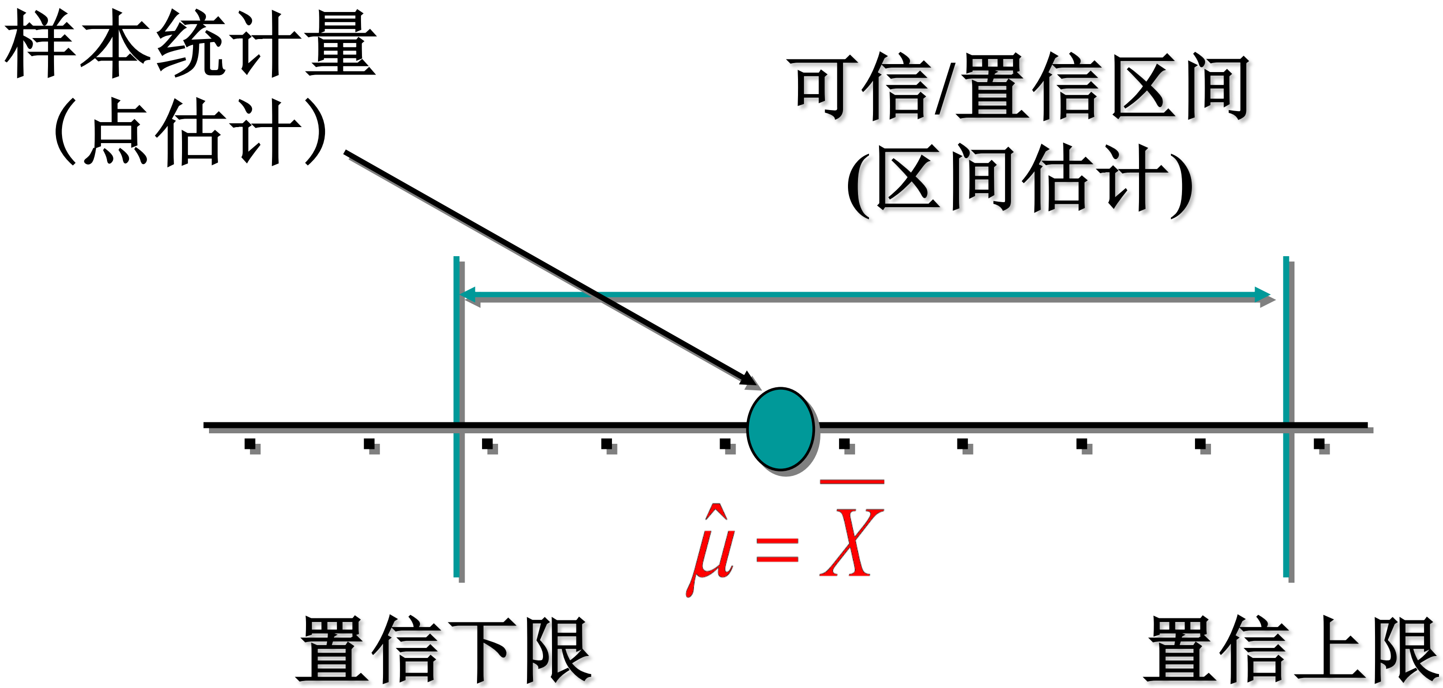
可信区间(confidence interval, CI)



总体均数的估计

点估计：**point estimation**

区间估计：**interval estimation**



置信区间/可信度

1. 重复试验时该区间包含总体均数的概率

2. 表示为 $1-\alpha$ 或 $100(1-\alpha)\%$

   常用的有 99%, 95%, 90%

   相应的 α 为0.01，0.05，0.10

置信区间/可信度

1. σ未知：

$$1-\alpha = P\left(-t_{\alpha/2,\nu} < \frac{\overline{X}-\mu}{S_{\overline{X}}} < t_{\alpha/2,\nu}\right)$$

$$\overline{X} - t_{\alpha/2,\nu} S_{\overline{X}} < \mu < \overline{X} + t_{\alpha/2,\nu} S_{\overline{X}}$$

总体均数μ的$(1-\alpha)$的可信区间为：

$(\bar{X} - t_{\alpha/2,\nu} S_{\bar{X}}, \ \bar{X} + t_{\alpha/2,\nu} S_{\bar{X}})$，或 $\bar{X} \pm t_{\alpha/2,\nu} S_{\bar{X}}$ 。

练习

$\overline{X}=5.03$，$S=0.52$，$n=10$

求总体均数的95%可信区间。

解：

由$n=10$，$\nu=10-1=9$，$\alpha=0.05$（双侧），

得$t_{0.05/2,9}=2.262$，

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

即 $:4.66<\mu<5.40$

总体均数得95%的可信区间为4.66～5.40（$10^{12}/L$）。

置信区间/可信度

2. σ已知，或σ未知但n足够大：

$$1-\alpha = P\left(-u_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma_{\overline{X}}} < u_{\alpha/2}\right)$$

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

σ未知，n足够大，故计算时用$\sigma_{\overline{X}}$的估计值$S_{\overline{X}}$

总体均数μ的$(1-\alpha)$的可信区间为：

$(\bar{X}-u_{\alpha/2}\sigma_{\bar{X}}$，$\bar{X}+u_{\alpha/2}\sigma_{\bar{X}})$，或 $\bar{X}\pm u_{\alpha/2}\sigma_{\bar{X}}$ 。

练 习

2003年随机抽取85位当地20岁应征男青年，其身高均数为171. 2 cm， 标准差为5. 3 cm。问该地区20岁男青年身高总体均数的95％的可信区间

解：

由 $n=85$，视为大样本数据，

$$\therefore 171.2 \pm 1.96 \times \frac{5.3}{\sqrt{85}} = (170.1,\ 172.3)\ (cm)$$

2003年当地20岁应征男青年身高总体均数的95%的可信区间为170.1～172.3cm。

95%可信区间的含义

从总体中作随机抽样，例如作100次抽样，每个样本可算得一个可信区间，得100个可信区间，平均有95个可信区间包括总体均数μ(估计正确)，只有5个可信区间不包括总体均数μ(估计不正确)。

实际中，只作一次抽样，只得到一个可信区间，作为未知总体均数的可能范围的估计，理论上有95％的可能是正确的，而5％的可能发生错误。