

第五章 参数估计

- 第一节 抽样分布
- 第二节 点估计
- 第三节 单样本总体均值的区间估计
- 第四节 两个独立样本总体均值差的区间估计
- 第五节 两个匹配样本总体均值差的区间估计
- 第六节 样本容量的确定
- 第七节 总体方差的区间估计
- 第八节 总体比率的区间估计
- 第九节 抽样组织形式

引例

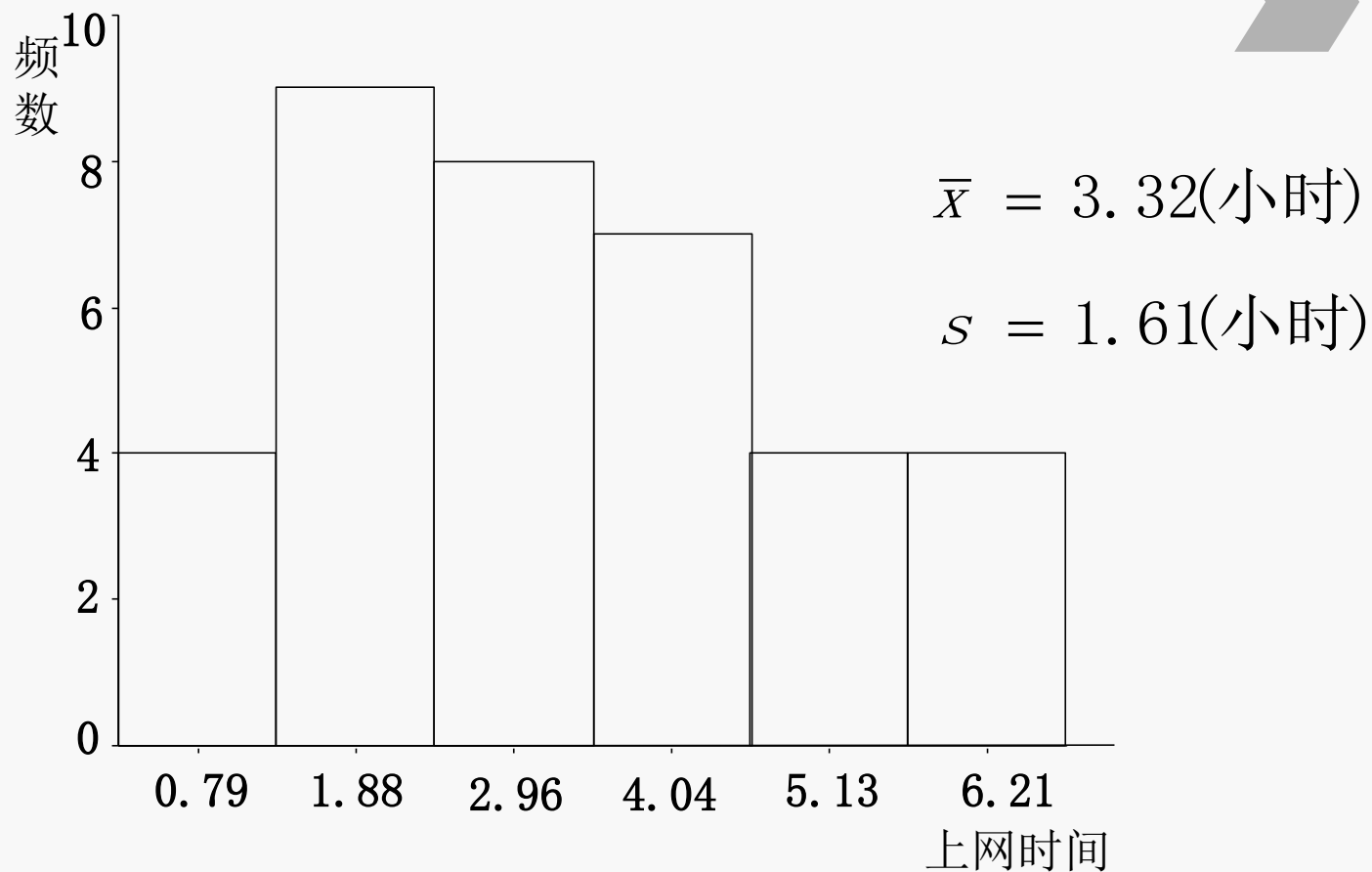
学校网管中心为合理制定校园网络管理条例，需要掌握每天全校学生的平均上网时间。从全校10000名学生中随机抽取了36名学生，调查他们每天的上网时间，获得样本数据：

36名学生每天上网时间样本数据（小时）

3.3	3.1	6.2	5.8	2.3	4.1
5.4	4.5	3.2	4.4	2.0	5.4
2.6	6.4	1.8	3.5	5.7	2.3
2.1	1.9	1.2	5.1	4.3	4.2
3.6	0.8	1.5	4.7	1.4	1.2
2.9	3.5	2.4	0.5	3.6	2.5



引例



36名学生每天上网时间的频数分布图



第一节 抽样分布

一、三种不同性质的分布

二、一个总体参数推断时样本统计量的抽样分布

三、两个总体参数推断时样本统计量的抽样分布



一、三种不同性质的分布

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

总体的分布往往是不知道的，通常是根据经验大致了解总体的分布类型，或者可以假定它服从某种参数，我们关心的是总体参数，如均值、比例、方差等(赌局游戏)。

- **样本分布 (sample distribution)**：如果从总体中抽取一个容量为 n 的样本，那么这 n 个观察值的分布就是样本分布。

样本是从总体中抽取的，其中包含着总体的一些信息和特征。当样本容量逐渐增大时，样本分布逐渐接近总体分布，但由于样本是随机抽取的，当样本容量很小时，样本分布就有可能与总体分布不一致。



一、三种不同性质的分布

- 总体分布 (**population distribution**)
- 样本分布 (**sample distribution**)
- 抽样分布 (**sampling distribution**)：样本统计量作为随机变量的概率分布。

统计量是一个随机变量，总体参数根据样本统计量来推断，这种推断的理论依据是样本统计量的抽样分布。

某个样本统计量的抽样分布，就是在重复选取容量为 n 的样本组时，由每一组样本算出的该统计量数值的相对频数分布。

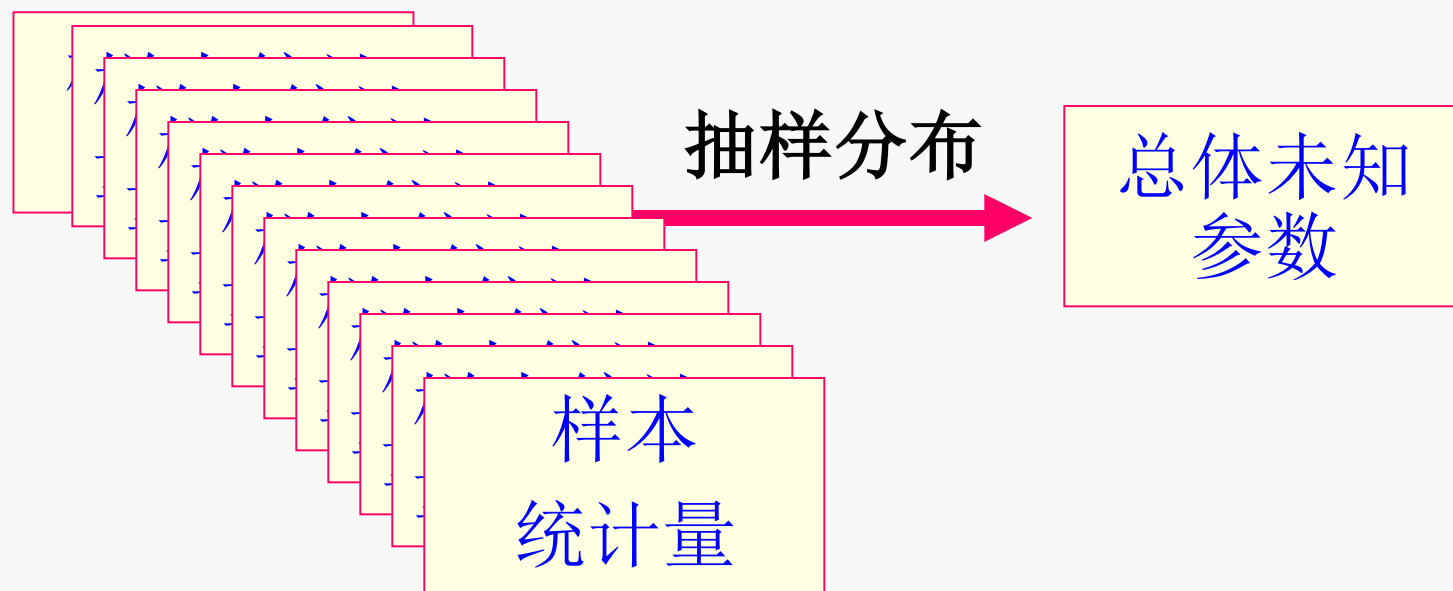
现实中我们不可能将所有可能的样本组都抽出来，因此统计量的抽样分布实际上是一种理论分布。



一、三种不同性质的分布

抽样分布

比如，用样本均值、比例、方差来推断总体的均值、比例、方差。



统计推断的过程示意图



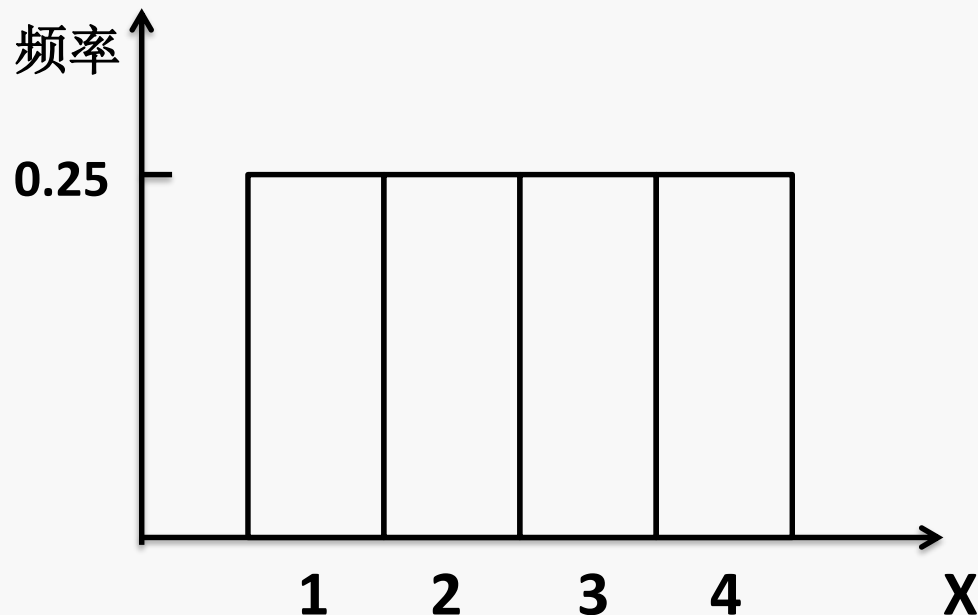
二、一个总体参数推断时样本统计量的抽样分布

(一) 样本均值的抽样分布

1. 样本均值抽样分布的形成过程

【例】 设一个总体有4个元素（个体），即总体元素个数 $N=4$ ，4个元素取值分别为1、2、3、4。

总体的分布状况为：



总体均值：2.5
总体方差：1.25



二、一个总体参数推断时样本统计量的抽样分布

从总体中采取重复抽样方法随机抽取 $n=2$ 的样本，共有 $4^2=16$ 个可能的样本。然后计算每一个样本的均值。



样本	样本中的元素	样本均值	样本方差
1	1,1	1.0	0
2	1,2	1.5	0.25
3	1,3	2.0	1
4	1,4	2.5	2.25
5	2,1	1.5	0.25
6	2,2	2.0	0
7	2,3	2.5	0.25
8	2,4	3.0	1
9	3,1	2.0	1
10	3,2	2.5	0.25
11	3,3	3.0	0
12	3,4	3.5	0.25
13	4,1	2.5	2.25
14	4,2	3.0	1
15	4,3	3.5	0.25
16	4,4	4.0	0

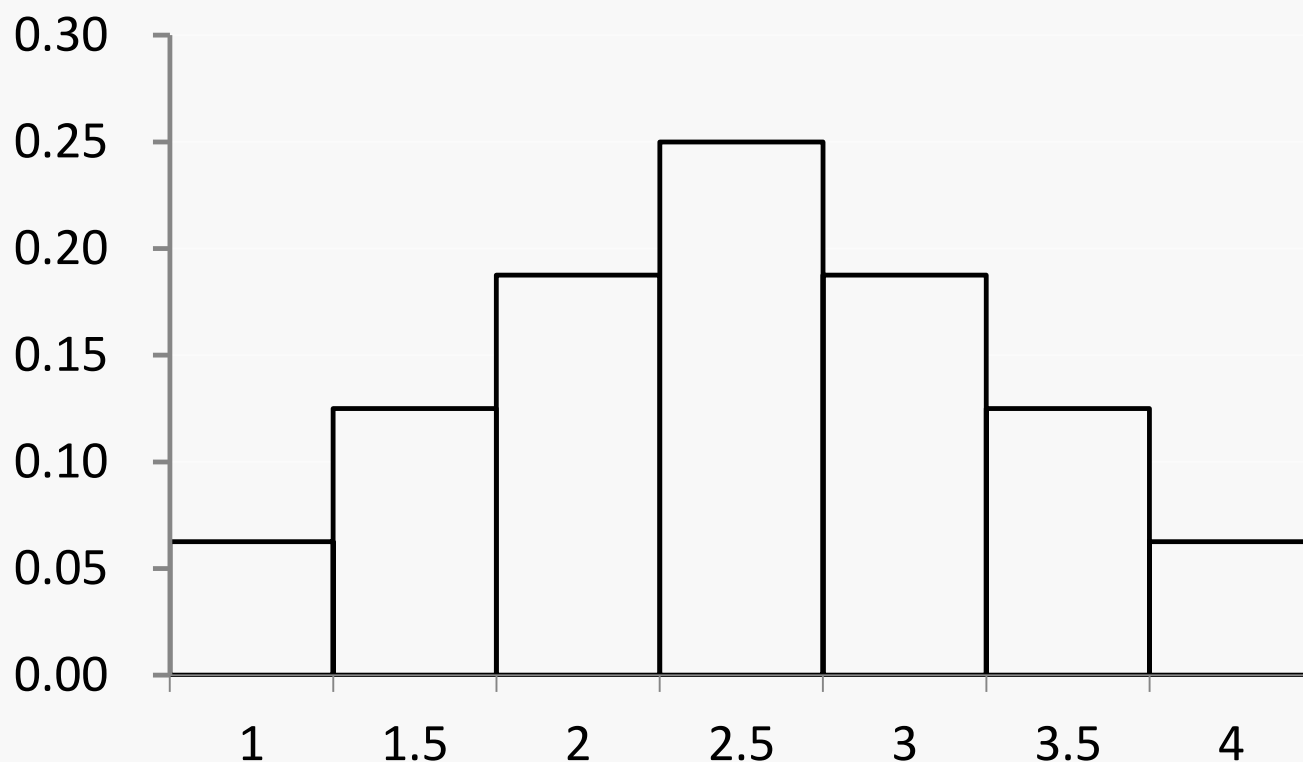
二、一个总体参数推断时样本统计量的抽样分布

样本均值的分布

均值的取值	均值的个数	均值取值的概率P
1.0	1	1/16
1.5	2	2/16
2.0	3	3/16
2.5	4	4/16
3.0	3	3/16
3.5	2	2/16
4.0	1	1/16



二、一个总体参数推断时样本统计量的抽样分布

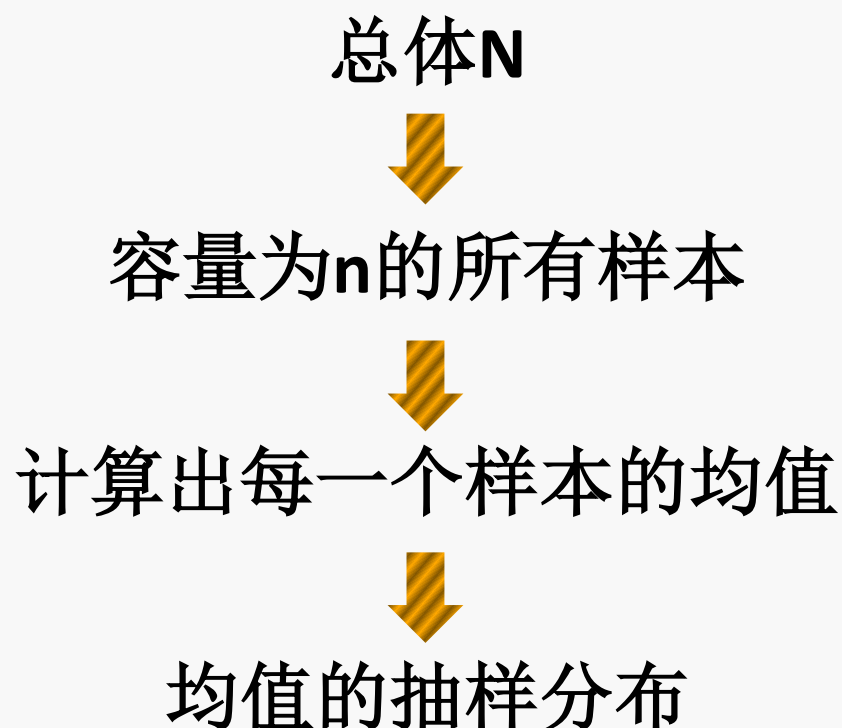


样本均值的抽样分布



二、一个总体参数推断时样本统计量的抽样分布

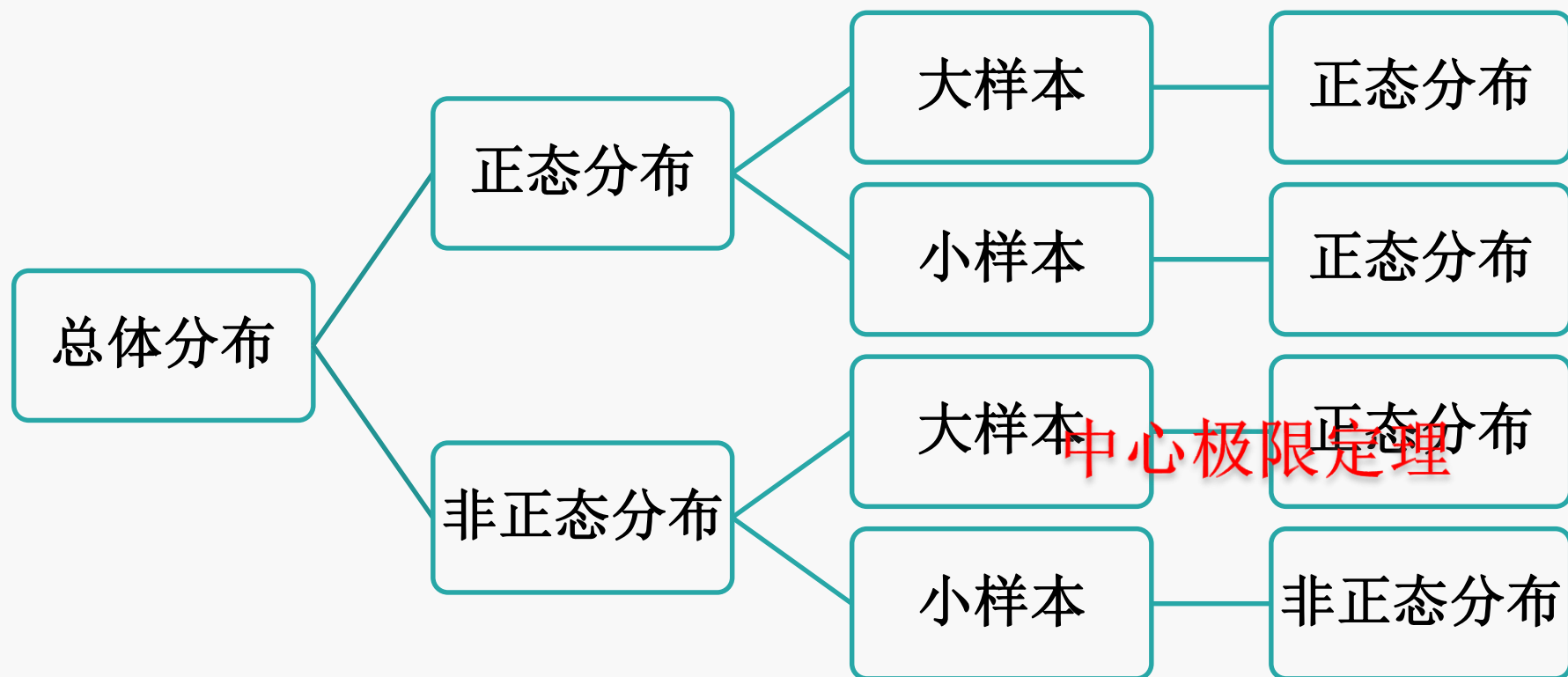
抽样分布(均值)的形成过程:



二、一个总体参数推断时样本统计量的抽样分布

2. 样本均值抽样分布的形式

样本均值抽样分布的形式与原有总体的分布和样本容量的大小有关。



二、一个总体参数推断时样本统计量的抽样分布

中心极限定理（central limit theorem）：从均值为 μ ，方差为 σ^2 的总体中，抽取容量为 n 的随机样本，当 n 充分大时（实际使用中通常要求 $n \geq 30$ ），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。

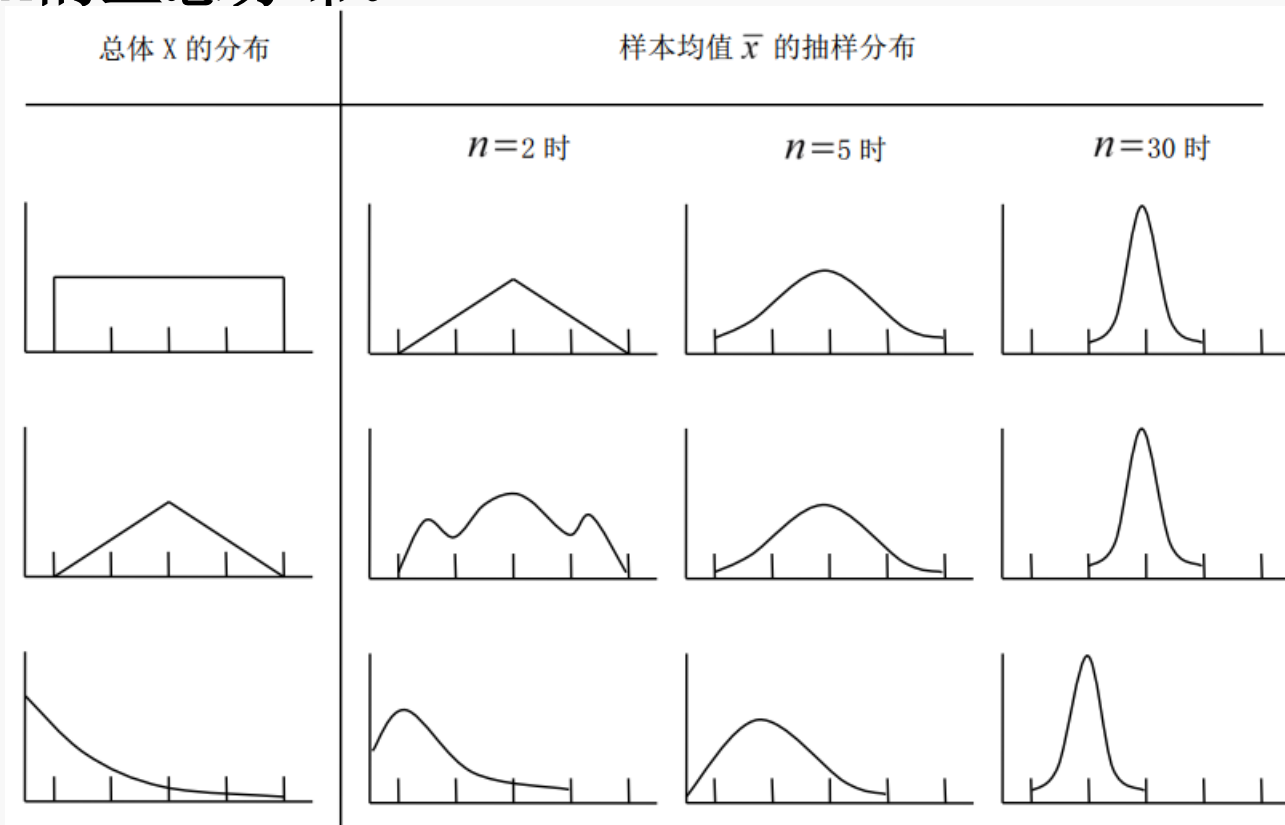


图 5-4 中心极限定理的作用

二、一个总体参数推断时样本统计量的抽样分布

3. 样本均值抽样分布的特征

从抽样推断的角度看，我们关心的分布特征主要是样本均值抽样分布的数学期望和方差，这不仅与**总体分布的均值和方差**有关，还与抽样的方法是**重复抽样还是不重复抽样**有关(抽取的样本是否独立)。



二、一个总体参数推断时样本统计量的抽样分布

设总体共有 N 个元素，均值为 μ ，方差为 σ^2 ，从中抽取容量为 n 的样本。

样本均值的数学期望为： $E(\bar{X}) = \mu$

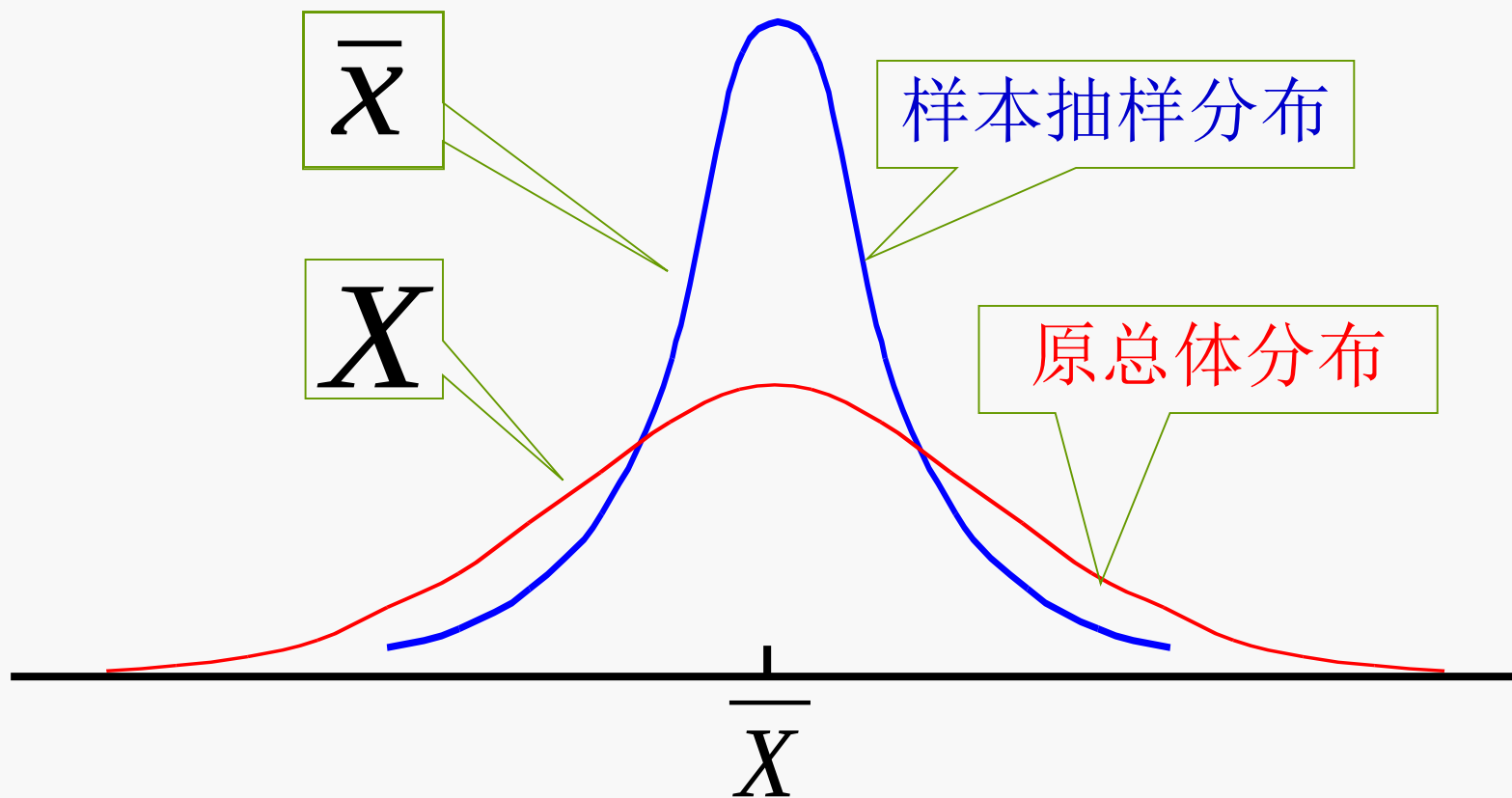
重复抽样下：

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

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

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





二、一个总体参数推断时样本统计量的抽样分布

(二) 样本比率的抽样分布

总体比率是指总体中具有某种属性的个体单位数与总体单位总数的比值，记作 P ；

样本比率是指样本中具有某种属性的样本单位数与样本单位总数的比值，记为 p 。

$P=N_1/N$ ， $p=n_1/n$ ，当 $np \geq 5$ ， $n(1-p) \geq 5$ ，且重复抽样条件下，有：

$$E(p) = P \qquad \sigma_p^2 = \frac{P(1-P)}{n}$$

$$p \sim N\left(P, \frac{P(1-P)}{n}\right), n \geq 30$$



二、一个总体参数推断时样本统计量的抽样分布

（三）样本方差的抽样分布

用样本方差去估计总体的方差，也必须知道样本方差的抽样分布。数理统计证明，对于来自正态总体的简单随机样本：

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

卡方分布：（1）变量值始终为正；（2）分布的形状取决于其自由度n的大小，通常为不对称的右偏分布，但随着自由度的增大逐渐趋于对称，最终趋于正态分布。

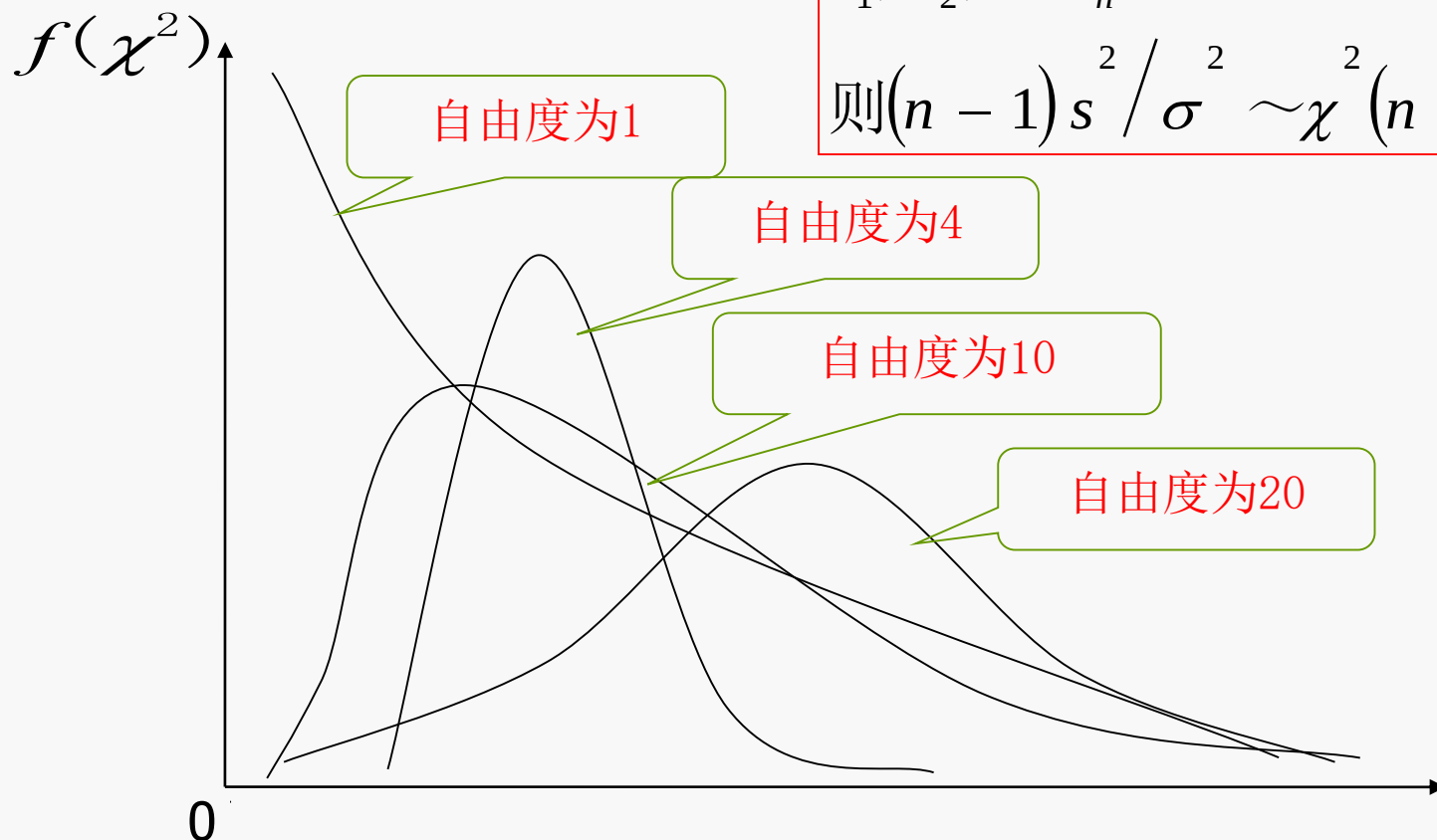


二、一个总体参数推断时样本统计量的抽样分布

若总体 $X \sim N\left(\mu, \sigma^2\right)$,

$x_1, x_2, \Lambda x_n$ 是取自总体的随机样本

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



$(n-1)s^2/\sigma^2$ 的抽样分布(χ^2 分布)

不同自由度下的卡方分布



第二节 点估计

一、点估计的含义

二、点估计量的优良标准



二、点估计量的优良标准

点估计

指用样本统计量的值直接作为对应总体参数的估计值。

点估计量的优良标准：

设 θ 为待估计的总体参数， $\hat{\theta}$ 为样本统计量，则 $\hat{\theta}$ 的优良标准为：

1. 无偏性 样本指标的均值应等于被估计的总体指标

若 $E(\hat{\theta}) = \theta$ ，则称 $\hat{\theta}$ 为 θ 的无偏估计量



二、点估计量的优良标准

点估计量的优良标准：

设 θ 为待估计的总体参数， $\hat{\theta}$ 为样本统计量，则 $\hat{\theta}$ 的优良标准为：

2. 有效性 作为同一个总体参数的两个无偏估计量，方差较小的更为有效

若 $\sigma_{\hat{\theta}_1} < \sigma_{\hat{\theta}_2}$ ，则称 $\hat{\theta}_1$ 为比 $\hat{\theta}_2$ 更有效的估计量

3. 一致性 指随着样本单位数 n 的增大，样本估计量将在概率意义下越来越接近于总体真实值

若 n 越大 $\sigma_{\hat{\theta}}$ 越小，则称 $\hat{\theta}$ 为 θ 的一致估计量



二、点估计量的优良标准

数理统计证明：

- $\bar{X}$ 为 μ 的无偏、更有效、一致估计量
- S_{n-1} 为 σ 的无偏、更有效、一致估计量
- p 为 P 的无偏、更有效、一致估计量

优点 简单，具体明确

缺点 无法控制误差，仅适用于对推断的准确程度与可靠程度要求不高的情况



第三节 单样本总体均值的区间估计

一、总体方差已知

二、总体方差未知

三、运用SPSS进行单样本总体均值的区间估计



区间估计

根据样本指标和抽样极限误差以一定的可靠程度推断总体指标的可能范围；被推断的总体指标的下限与上限所包括的区间称为置信区间，估计的可靠程度称为置信度。



一、总体方差已知

由抽样分布可知：

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

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

➤ 给定区间 $[-Z_{\alpha/2}, Z_{\alpha/2}]$ 使得

$$P(Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1-\alpha$$



一、总体方差已知

总体均值的区间估计：

计算公式

$$\bar{x} - \Delta_{\bar{x}} \leq \bar{X} \leq \bar{x} + \Delta_{\bar{x}},$$

$$\text{或 } \bar{X} \in [\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}]$$

式中, $\Delta_{\bar{x}} = Z_{\alpha/2} \mu_{\bar{x}}$

- $\Delta_{\bar{x}}$ 为抽样极限误差，指在一定的概率保证程度下，抽样误差不允许超过的某一给定范围，也称作允许误差、误差范围、误差置信限等
- $Z_{\alpha/2}$ 为概率度，是给定概率保证程度下样本均值偏离总体均值的抽样平均误差的倍数



一、总体方差已知

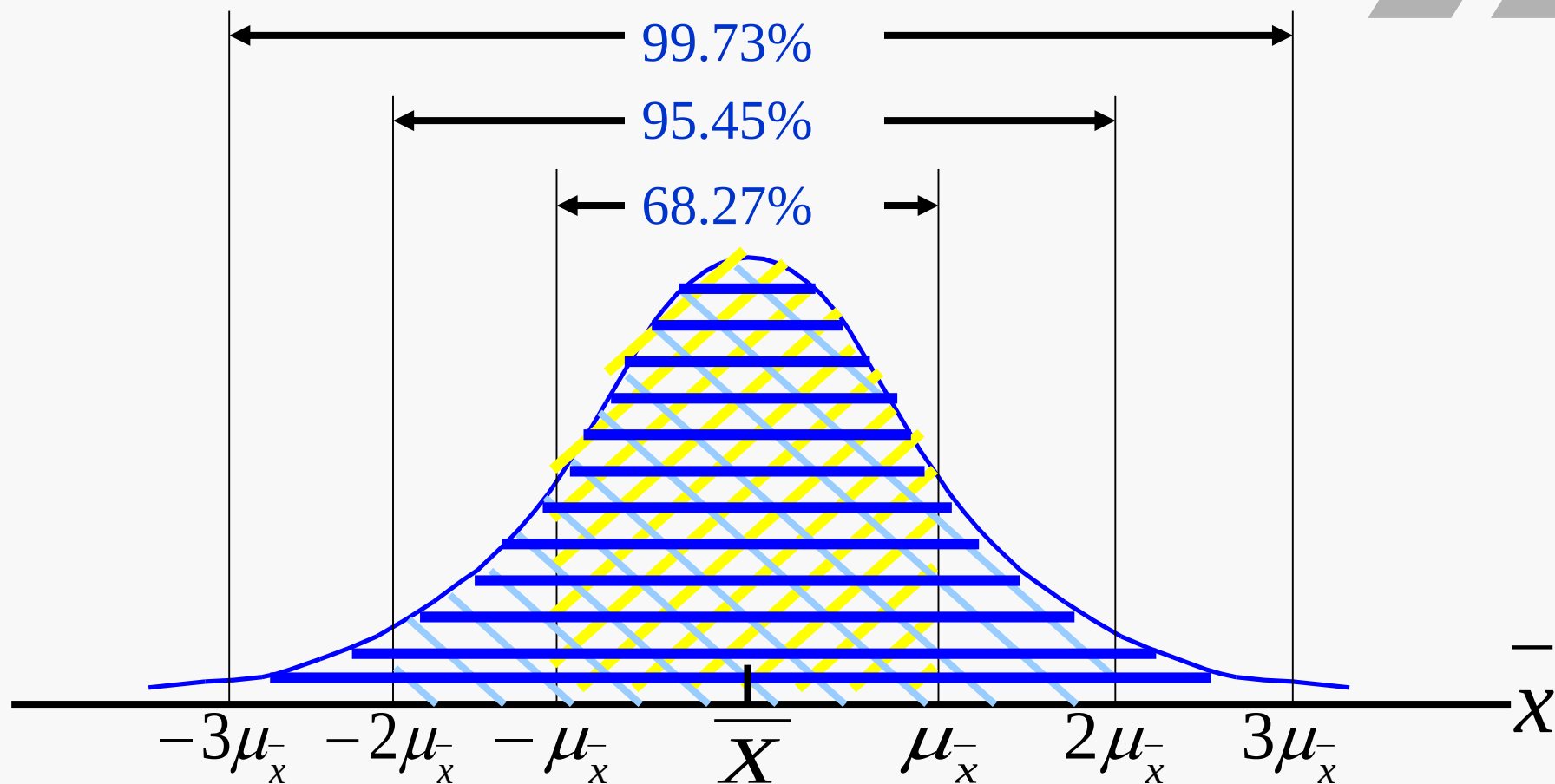
- 提高把握程度，会增大允许误差，使估计精度降低，而缩小允许误差，提高估计的精度，又会降低估计的把握程度。在实际中应根据具体情况，先确定一个合理的把握程度，再求相应的允许误差，或先确定一个允许误差范围，再求相应的把握程度。 $Z_{\alpha/2}$ 与相应的概率保证程度存在一一对应关系。

常用 $Z_{\alpha/2}$ 值及相应的
概率保证程度

$Z_{\alpha/2}$ 值	概率保证程度
1.00	0.6827
1.65	0.9000
1.96	0.9500
2.00	0.9545
2.58	0.9900
3.00	0.9973

$1 - \alpha$





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



一、总体方差已知

步骤:

1. 计算样本平均数 $\bar{x}$;

2. 计算抽样平均误差: $\mu_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

3. 计算抽样极限误差: $\Delta_{\bar{x}} = Z_{\alpha/2} \mu_{\bar{x}}$

4. 确定总体平均数的置信区间:

$$\bar{x} - \Delta_{\bar{x}} \leq \bar{X} \leq \bar{x} + \Delta_{\bar{x}},$$

$$\text{或 } \bar{X} \in [\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}]$$



一、总体方差已知



【例5.1】从袋装食品生产线上随机抽取25袋测量其重量，算得样本均值为105.36克。已知该生产线上所生产袋装食品总体的重量服从正态分布，总体标准差为10克，若给定置信度为95%，试求总体均值的估计区间。



一、总体方差已知

解：

$\bar{x} \sim N(\mu, \sigma^2/n)$ ，根据公式有：

总体均值的估计区间为

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 105.36 \pm 1.96 \times \frac{10}{\sqrt{25}} = 105.36 \pm 3.92$$

即[101.44, 109.28]。结果表明，该批袋装食品的平均重量在101.44克到109.28克之间，做出这一推断的把握程度为95%。



二、总体方差未知

总体均值的区间估计：

计算公式：

$$\bar{x} - \Delta_x \leq \bar{X} \leq \bar{x} + \Delta_x ,$$
$$\text{或 } \bar{X} \in [\bar{x} - \Delta_x , \bar{x} + \Delta_x]$$

式中， $\Delta_x = Z_{\alpha/2} \mu_x$ 为极限误差



二、总体方差未知

步骤:

1. 计算样本平均数 $\bar{x}$;
2. 搜集总体方差的经验数据 σ^2 ; 或计算样本方差 s^2 , 即

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}, s^2 = \frac{\sum (x - \bar{x})^2 f}{\sum f - 1}$$

3. 计算抽样平均误差: $\mu_{\bar{x}} = \frac{s}{\sqrt{n}}$



二、总体方差未知

步骤:

4. 计算抽样极限误差: $\Delta_{\bar{x}} = t_{\alpha/2} \mu_{\bar{x}}$

5. 确定总体平均数的置信区间:

$$\bar{x} - \Delta_{\bar{x}} \leq \bar{X} \leq \bar{x} + \Delta_{\bar{x}},$$
$$\text{或 } \bar{X} \in [\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}]$$



二、总体方差未知

【例5.2】 本章引例中，网管中心抽取了一个容量为36的样本，并计算得出： $\bar{x} = 3.32$ （小时）， $s = 1.61$ （小时）。若给定置信度为95%，试求每天全校学生平均上网时间的估计区间。



二、总体方差未知

解：

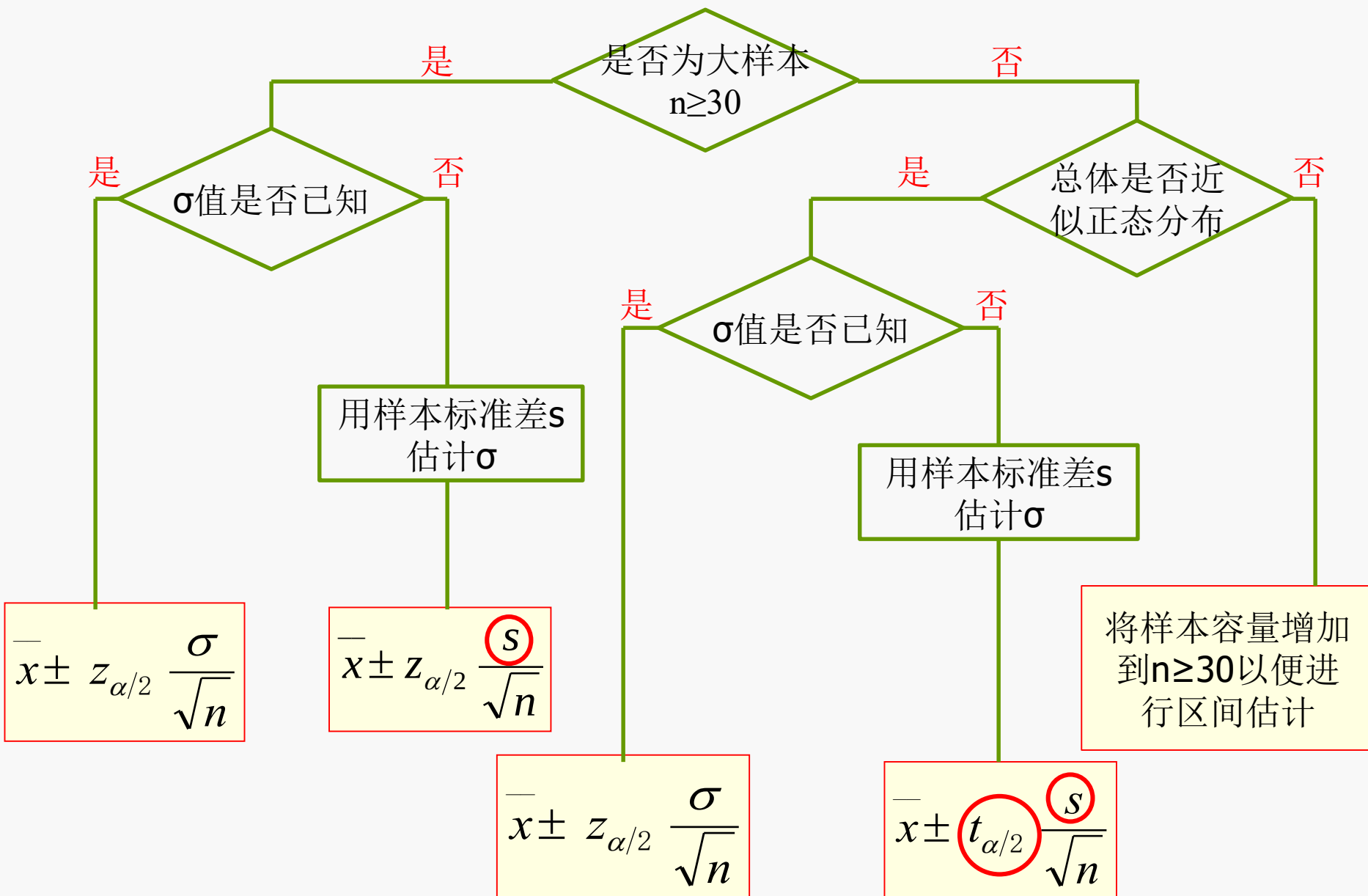
根据公式有：

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 3.32 \pm 2.03 \times \frac{1.61}{\sqrt{36}} = 3.32 \pm 0.54$$

即[2.78, 3.86]。结果表明，每天全校学生平均上网时间在2.78小时到3.86小时之间，做出这一推断的把握程度为95%。



总体均值区间估计程序



三、运用SPSS进行单样本总体均值的区间估计

操作步骤：

- 打开数据，选择检验方法：

Analyze→Compare Means→One Sample T Test

- 设置变量：

“time” →Test Variable(s)

- 设置检验值：

Test Value: 0

- 设置检验置信度：

Options→Confidence interval→Continue→OK



三、运用SPSS进行单样本总体均值的区间估计

输出结果：

One-Sample Test						
	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
上网时间	12.365	35	.000	3.3166667	2.772142	3.861192

即可以据此得出结论：在95%的置信度水平下，该高校学生每天上网时间在2.78到3.86小时之间。



第四节 两个独立样本总体均值差的区间估计

- 一、大样本，方差已知
- 二、大样本，方差未知
- 三、小样本，方差未知，但方差相等
- 四、小样本，方差未知，方差不相等
- 五、运用SPSS进行两个独立样本总体均值差的区间估计



理论基础:

- 数理统计的研究表明, 分别由两个正态总体独立抽取任意容量的样本, 或者由两个任意总体独立抽取容量充分大的样本, 则两个样本的均值差 $\bar{x}_1 - \bar{x}_2$ 服从以 $\mu_1 - \mu_2$ 为数学期望, 以 $\sigma_1^2/n_1 + \sigma_2^2/n_2$ 为方差的正态分布。



一、大样本，方差已知

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

➡ $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ (构造Z统计量)

➡ $(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ (估计区间)



一、大样本，方差已知



【例5.3】 随机独立抽取男女中学生各30人测量体重，获得两个样本数据，并计算出男女中学生样本体重的均值分别为 $\bar{x}_1 = 48.74$, $\bar{x}_2 = 43.05$ 。已知男女中学生体重的总体方差分别为 $\sigma_1^2 = 47.78$, $\sigma_2^2 = 20.93$ 。若给定置信度为95%，试求男女中学生总体平均体重之差的估计区间。



一、大样本，方差已知

解：

$$\begin{aligned} \text{根据公式有 } & (\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= (48.74 - 43.05) \pm 1.96 \times \sqrt{\frac{47.78}{30} + \frac{20.93}{30}} \\ &= 5.69 \pm 1.96 \times 1.51 \\ &= 5.69 \pm 2.96 \end{aligned}$$

即：[2.73, 8.65]。结果表明，男女中学生体重之差平均在2.73公斤到8.65公斤之间，做出这一推断的把握程度为95%。



二、大样本，方差未知

分别以两个样本方差 s_1^2 和 s_2^2 作为两个总体方差 σ_1^2 和 σ_2^2 的点估计，并进一步构造 **t 统计量**：


$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$


$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (\text{估计区间})$$

(式中： $Z_{\alpha/2}$ 替代了 $t_{\alpha/2}$ ， t 分布自由度较大时，接近标准正态)



二、大样本，方差未知

【例5.4】随机独立抽取男女中学生各30人测量身高，获得两个样本数据，并计算出男女中学生样本体重的均值和样本方差分别为 $\bar{x}_1 = 162.67$, $\bar{x}_2 = 158.59$, $s_1^2 = 68.52$, $s_2^2 = 41.77$ 。若男女中学生身高总体的方差未知，试求给定置信度为95%时，男女中学生总体平均身高之差的估计区间。



二、大样本，方差未知

解：

$$\begin{aligned} \text{根据公式有 } & (\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ & = (162.67 - 158.59) \pm 1.96 \times \sqrt{\frac{68.52}{30} + \frac{41.77}{30}} \\ & = 4.08 \pm 1.96 \times 1.92 \\ & = 4.08 \pm 3.76 \end{aligned}$$

即：[0.32, 7.84]。结果表明，男女中学生身高之差平均在0.32厘米到7.84厘米之间，做出这一推断的把握程度为95%。



三、小样本，方差未知，但方差相等

构造自由度为 $(n_1 + n_2 - 2)$ 的 t 统计量：


$$t = \frac{\left(\bar{x}_1 - \bar{x}_2 \right) - \left(\mu_1 - \mu_2 \right)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$


$$\left(\bar{x}_1 - \bar{x}_2 \right) \pm t_{\alpha/2} (n_1 + n_2 - 2) \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

(估计区间)



三、小样本，方差未知，但方差相等

【例5.5】 分别在加工同一种零件的两条生产线上，各随机抽取12个零件并测量长度，获得两个样本数据，计算出的样本均值和样本方差分别为 $\bar{x}_1 = 32.5$, $\bar{x}_2 = 28.8$, $s_1^2 = 15.10$, $s_2^2 = 19.36$ 。假定两条生产线所生产零件的长度均服从正态分布，而且方差相等，若给定置信度95%，试求两条生产线所生产零件的平均长度之差的区间估计。



三、小样本，方差未知，但方差相等

解：

根据公式有

$$\begin{aligned} & (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}(n_1 + n_2 - 2) \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \\ &= (32.5 - 28.8) \pm t_{0.025}(22) \sqrt{\left(\frac{1}{12} + \frac{1}{12}\right) \frac{11 \times 15.10 + 11 \times 19.36}{12 + 12 - 2}} \\ &= 3.70 \pm 2.074 \times \sqrt{0.17 \times 17.23} \\ &= 3.70 \pm 3.55 \end{aligned}$$

即：[0.15, 7.25]。结果表明，两条生产线所生产零件的长度之差平均在0.15厘米到7.25厘米之间，做出这一推断的把握程度为95%。



四、小样本，方差未知，方差不相等

两个总体均值差 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 估计区间为：

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}(v) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

式中： $v = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 / \left[\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1} \right]$ 为 t 统计量的自由度。



四、小样本，方差未知，方差不相等

【例5.6】为确认同一种设备的两种组装工艺的优劣，随机抽取10名工人，采用第一种工艺进行组装操作，并记录每名工人所需的时间；再随机抽取另外10名工人，采用第二种工艺进行组装操作，并记录每名工人所需的时间。经计算得出两个样本的均值及方差分别为 $\bar{x}_1 = 31.61$, $\bar{x}_2 = 27.72$, $s_1^2 = 14.16$, $s_2^2 = 22.16$. 假定两种组装工艺所需时间均服从正态分布，但方差不等。若给定置信度为95%，试估计两种组装工艺平均所需时间之差的区间估计。



四、小样本，方差未知，方差不相等

解：先计算自由度

$$\begin{aligned}v &= \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 / \left[\frac{(s_1^2 / n_1)^2}{n_1 - 1} + \frac{(s_2^2 / n_2)^2}{n_2 - 1} \right] \\&= \left(\frac{14.16}{10} + \frac{22.16}{10} \right)^2 / \left[\frac{(14.16 / 10)^2}{10 - 1} + \frac{(22.16 / 10)^2}{10 - 1} \right] \\&\approx 17\end{aligned}$$

根据公式有

$$\begin{aligned}&(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}(v) \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)} \\&= (31.61 - 27.72) \pm t_{0.025}(17) \sqrt{\left(\frac{14.16}{10} + \frac{22.16}{10} \right)} \\&= 3.89 \pm 4.03\end{aligned}$$

即： $[-0.14, 7.92]$ 。结果表明，两种组装工艺所需时间之差平均在-0.14到7.92之间，做出这一推断的把握程度为95%。



五、运用SPSS进行两个独立样本总体均值差的区间估计

操作步骤:

- 打开数据，选择检验方法:

Analyze→Compare Means→Independent-Samples T Test

- 设置变量:

“score” →Test Variable(s)

“school” →Grouping Variable

- 定义分组变量:

Use specified values→Continue

- 设置检验置信度:

Options→Confidence interval→ Continue→OK



五、运用SPSS进行两个独立样本总体均值差的区间估计

输出结果：

Independent Samples Test			
		数学成绩	
		Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F	.029	
	Sig.	.868	
t-test for Equality of Means	t	-.544	-.544
	df	18	17.996
	Sig. (2-tailed)	.593	.593
	Mean Difference	-3.400	-3.400
	Std. Error Difference	6.249	6.249
	95% Confidence Interval of the Difference		
	Lower	-16.528	-16.529
	Upper	9.728	9.729

在假定两个独立样本的方差相等（即两个学院学生数学成绩方差相等）的情况下，两个学院学生数学成绩均值差的95%估计区间为[-16.528， 9.728]。

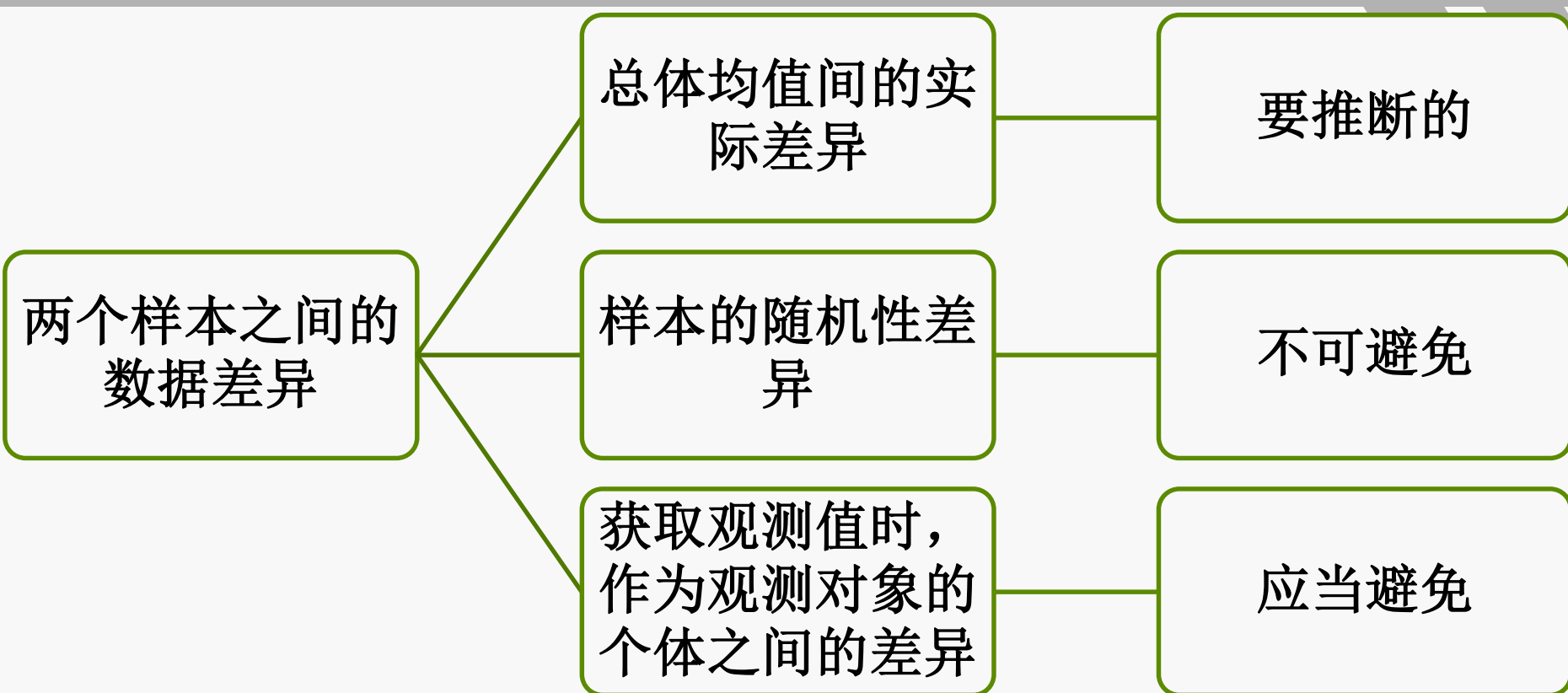


第五节 两个匹配样本总体均值差的区间估计

一、两总体均值差的区间估计

二、运用SPSS进行两总体均值差的区间估计





在独立样本的抽样方案中，并没有考虑到第三种差异来源的存在，因此在获取观测值时，如果作为观测对象的个体之间存在较大的差异，必然会降低估计的准确性。匹配样本就是试图避免第三种差异来源。



- 
- 例如，某工厂有两种方法供员工执行某项生产任务，考察工人的完成时间。第一个方案是：从生产工人中随机抽取一个简单样本，每名被抽中的工人使用第一种方法进行生产；再随机抽取另一个简单样本，每名被抽中的工人使用第二种方法进行生产。第二个方案是：从生产工人中随机抽取一个简单样本，每名被抽中的工人先用其中一种方法进行生产，然后再改用另一种方法进行生产。
 - 第一种方案得到的是独立样本，而第二种方案得到匹配样本。



- 匹配样本数据中，每一个观测对象都提供了一对观测值，在进行总体均值差区间估计时，需要对这种配对观测值进行特别的处理。首先要计算每一对观测值之间的差值 d ，再以 d 为被平均的对象，计算其均值 $\bar{d}$ 。



一、两总体均值差的区间估计

两总体间差值的方差已知：

由 $\bar{d} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_d^2}{n}\right)$

➡ $Z = \frac{\bar{d} - \mu_d}{\sigma_d / \sqrt{n}}$ (构造Z统计量)

➡ $\bar{d} \pm Z_{\alpha/2} \frac{\sigma_d}{\sqrt{n}}$ (估计区间)



一、两总体均值差的区间估计

两总体间差值的方差未知：

构造自由度为 $(n - 1)$ 的 t 统计量：

$$\Rightarrow t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

$$\Rightarrow \bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}} \quad (\text{估计区间})$$



二、运用SPSS进行两总体均值差的区间估计

操作步骤:

- 打开数据，选择检验方法:

Analyze→Compare Means→Paired-Sample T Test

- 设置变量:

“method1” →Variable1

“method2” →Variable2

- 设置检验置信度:

Options→Confidence interval→Continue→OK



二、运用SPSS进行两总体均值差的区间估计

输出结果：

Paired Samples Test			
		Pair 1	
		方法1的完成 时间 - 方法2的 完成时间	
Paired Differences	Mean	.3000	
	Std. Deviation	.3347	
	Std. Error Mean	.1366	
	95% Confidence Interval of the Difference	Lower	-.0512
		Upper	.6512
t	2.196		
df	5		
Sig. (2-tailed)	.080		

结果显示：在95%的置信度下，两种方法完成工作时间均值差的估计区间为[-0.0512， 0.6512]。



第六节 样本容量的确定

- 一、估计单个总体均值时样本容量的确定
- 二、估计总体均值差时样本容量的确定
- 三、样本容量的现实矫正



区间估计对总体参数的真值，给出了具有一定把握程度和准确程度的推断。

- 把握程度

体现在置信度 $1-\alpha$ 上，置信度越高，则估计的把握程度越高；相应地，估计区间约宽。

- 准确程度

体现在估计区间的宽窄上，估计区间越窄，则准确程度越高；相应地，置信度越低。

提高把握程度与提高准确程度是相互矛盾的，唯一的途径是扩大样本容量，但抽取样本要花费成本。

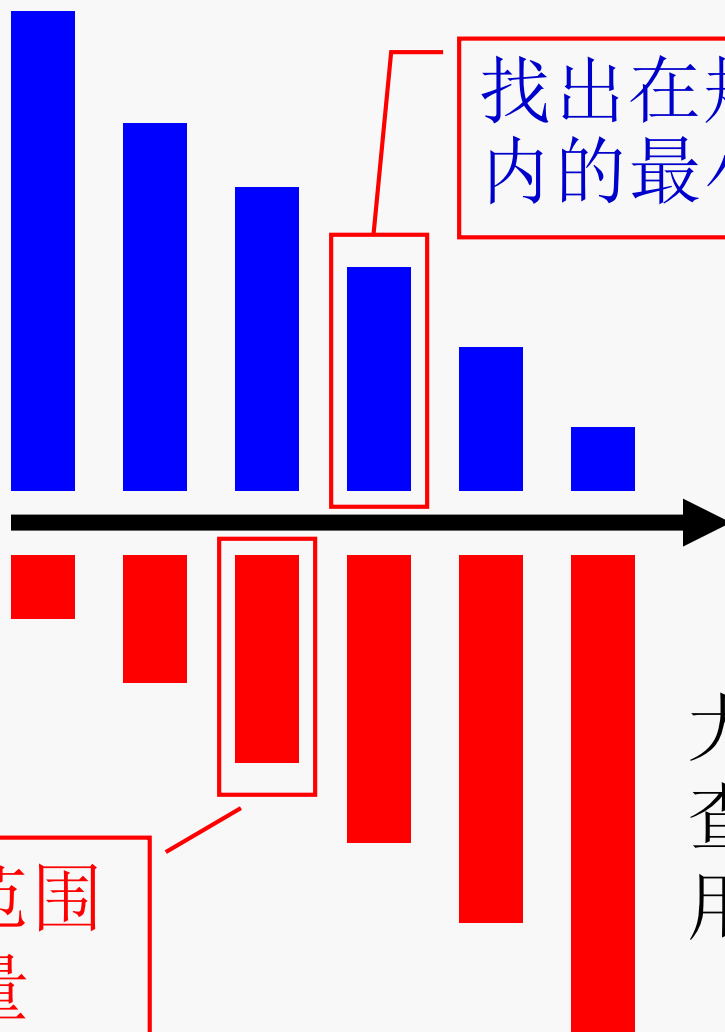


小样本容量节省费用，但调查误差大

调查误差

样本容量

调查费用



找出在规定误差范围内的最小样本容量

找出在限定费用范围内的最大样本容量

大样本容量调查精度高但费用较大



一、估计单个总体均值时样本容量的确定

通常的做法是先确定置信度，然后限定抽样极限误差。极限误差的大小决定了估计区间的宽窄，即估计的准确程度高低。

σ 或 S 通常未知。一般按以下方法确定其估计值：
①过去的经验数据；②
试验调查样本的 S 。

$$\ominus \Delta_{\bar{x}} = Z_{\alpha/2} \mu_x^- = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$
$$\therefore n = \frac{Z_{\alpha/2}^2 \sigma^2}{\Delta_{\bar{x}}^2} = \frac{\sigma^2}{\mu_x^-^2}$$

宁多勿少，计算结果通常向上进位



一、估计单个总体均值时样本容量的确定

【例5.8】 欲对某个调查总体的均值构造一个置信度为95%的估计区间，试调查所得到的样本的标准差为70。如果给定极限误差为20，试求所需的样本容量最小应该是多少？



解：

已知 $\sigma = 70$, $\Delta_{\bar{x}} = 20$, $Z_{\alpha/2} = 1.96$,

根据公式有：

$$n = \frac{Z_{\alpha/2}^2 \sigma^2}{\Delta_{\bar{x}}^2} = \frac{1.96^2 \times 70^2}{20^2} = 47.06$$

即所需的样本量应为48个。



二、估计两个总体均值差时样本容量的确定

计算公式：

$$n_1 = n_2 = \frac{(Z_{\alpha/2})^2 (\sigma_1^2 + \sigma_2^2)}{\Delta_x^2}$$

【例5.9】已知统计学院与经济学院学生往届高等数学成绩的标准差为：统计学院 $\sigma_1^2 = 100$ ，经济学院 $\sigma_2^2 = 120$ 。如果置信度为95%，极限误差为4，为比较本届两个学院学生高等数学成绩的差异情况，两个学院应当分别抽取多少样本容量。



二、估计两个总体均值差时样本容量的确定

解：两个学院应当分别抽取的样本容量为：

$$\begin{aligned} n_1 = n_2 &= \frac{\left(Z_{\alpha/2} \right)^2 \left(\sigma_1^2 + \sigma_2^2 \right)}{\Delta^2} \\ &= \frac{1.96^2 \times (100 + 120)}{4^2} = 52.82 \end{aligned}$$

即两个学院至少应当分别抽取容量为53的样本。



三、样本容量的现实矫正

- 为了达到理论样本量，必须接触更多的潜在调查对象，即需要在考虑发生率和完成率的条件下，保障调查的保证程度和准确程度
- 发生率 合格样本的样本框往往没有那么严格，因此需要发生率来矫正
- 比如，假定目标是25-55岁的女性，那么实际接触的群体可能是20-60岁的女性，这样有75%(发生率)是我们想要的。平均来讲，想获得一个合格的样本，我们得去接触1.33位20-60岁的女性。



三、样本容量的现实矫正

- 完成率 抽选到的合格对象也可能在调查执行过程中被拒绝，因此需要完成率来矫正。
- 假如希望符合这个条件的调查队形实现80%的访谈完成率，则实际接触的合格对象应该乘以1.25(1/0.8)这个因子。
- 将发生率和合格率结合起来即可确定现实所需要的样本量，比如理论样本量为300，则现实样本量应为 $300 \times 1.33 \times 1.25 = 500$.



第七节 总体方差的区间估计

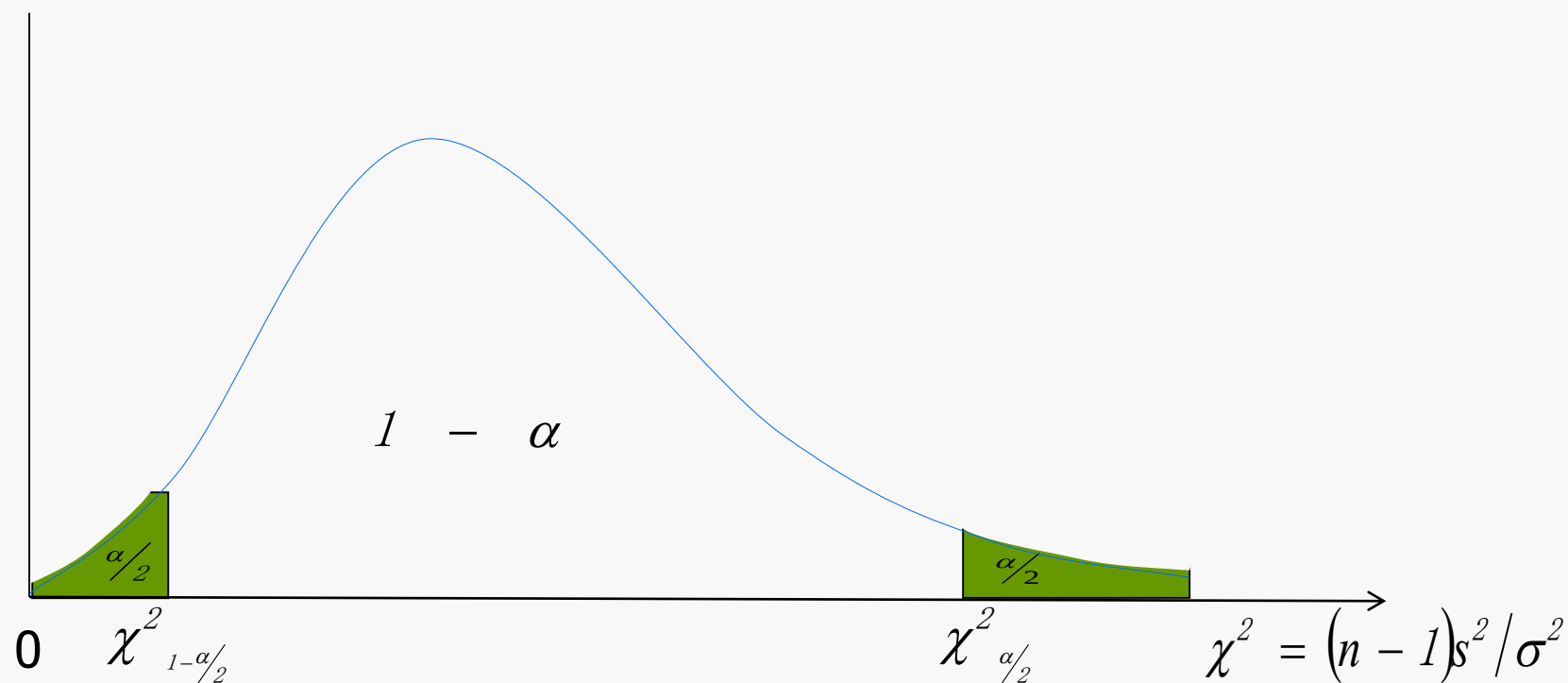
一、单样本总体方差的区间估计

二、两个独立样本总体均值差比的区间估计



一、单样本总体方差的区间估计

对于来自正态总体容量为 n 的简单随机样本，统计量 $(n-1)s^2/\sigma^2$ 服从自由度为 $(n-1)$ 的卡方分布。



一、单样本总体方差的区间估计

由
$$P\left(\chi_{1-\alpha/2}^2 \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{\alpha/2}^2\right) = 1 - \alpha$$


$$P\left(\frac{(n-1)s^2}{\chi_{\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}\right) = 1 - \alpha$$

即单样本总体方差 σ^2 在置信度 $1-\alpha$ 下的存在区间公式为：

$$\left(\frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)} \right)$$



一、单样本总体方差的区间估计

【例5.10】 在一条向塑料容器中灌装液体洗涤剂的生产线上，随机抽取了20罐，计算得每罐洗涤剂重量的样本方差为0.0025。试以95%的置信度，估计总体方差的存在区间。

$$(\chi_{0.025}^2(19) = 32.8523, \chi_{0.975}^2(19) = 8.90655)$$



一、单样本总体方差的区间估计

解： 总体方差 σ^2 的95%的估计区间下限为：

$$\frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)} = \frac{(20-1) \times 0.0025}{\chi_{0.025}^2(20-1)} = \frac{19 \times 0.0025}{\chi_{0.025}^2(19)} = \frac{0.0475}{32.8523} = 0.0014$$

总体方差 σ^2 的95%的估计区间上限为：

$$\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)} = \frac{(20-1) \times 0.0025}{\chi_{0.975}^2(20-1)} = \frac{19 \times 0.0025}{\chi_{0.975}^2(19)} = \frac{0.0475}{8.90655} = 0.0053$$

可以推断总体方差在0.0014至0.0053之间，做出这一推断的把握程度为95%。



二、两个独立样本总体均值差比的区间估计

- 根据样本方差的抽样分布原理可知，存在统计量 $(n_1 - 1)s_1^2 / \sigma_1^2$ 和 $(n_2 - 1)s_2^2 / \sigma_2^2$ ，分别服从自由度为 $(n_1 - 1)$ 和 $(n_2 - 1)$ 的卡方分布
- 数理统计相关理论证明，除以各自自由度的两个独立卡方变量之比，服从F分布，且F分布的两个自由度与两个卡方分布的自由度是相同的



二、两个独立样本总体均值方差比的区间估计

于是有 **F**统计量：

$$\frac{\frac{(n_1 - 1)s_1^2 / \sigma_1^2}{(n_1 - 1)}}{\frac{(n_2 - 1)s_2^2 / \sigma_2^2}{(n_2 - 1)}}$$

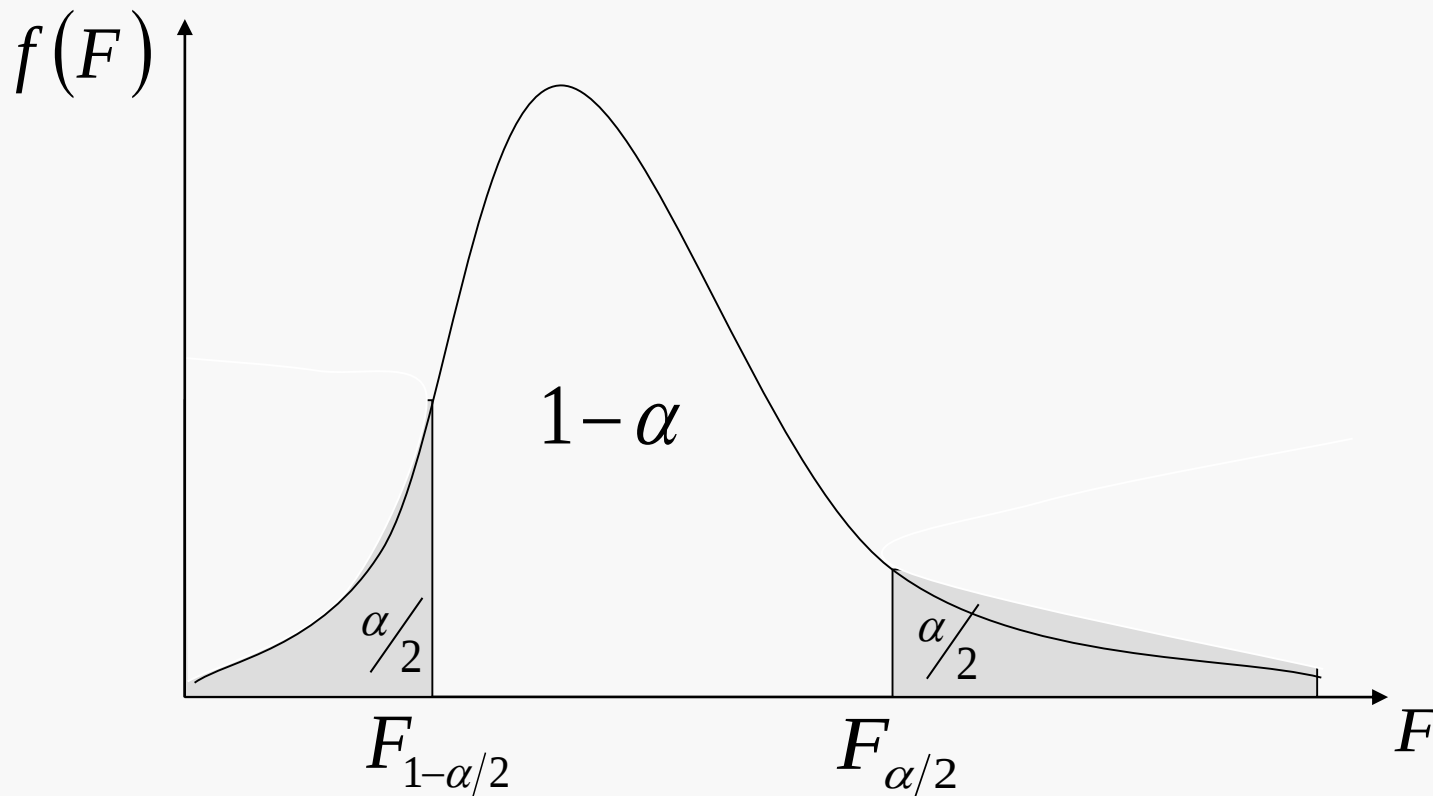
化简得：

$$\frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

服从分子自由度为 $n_1 - 1$ 、分母自由度为 $n_2 - 1$ 的
F分布。



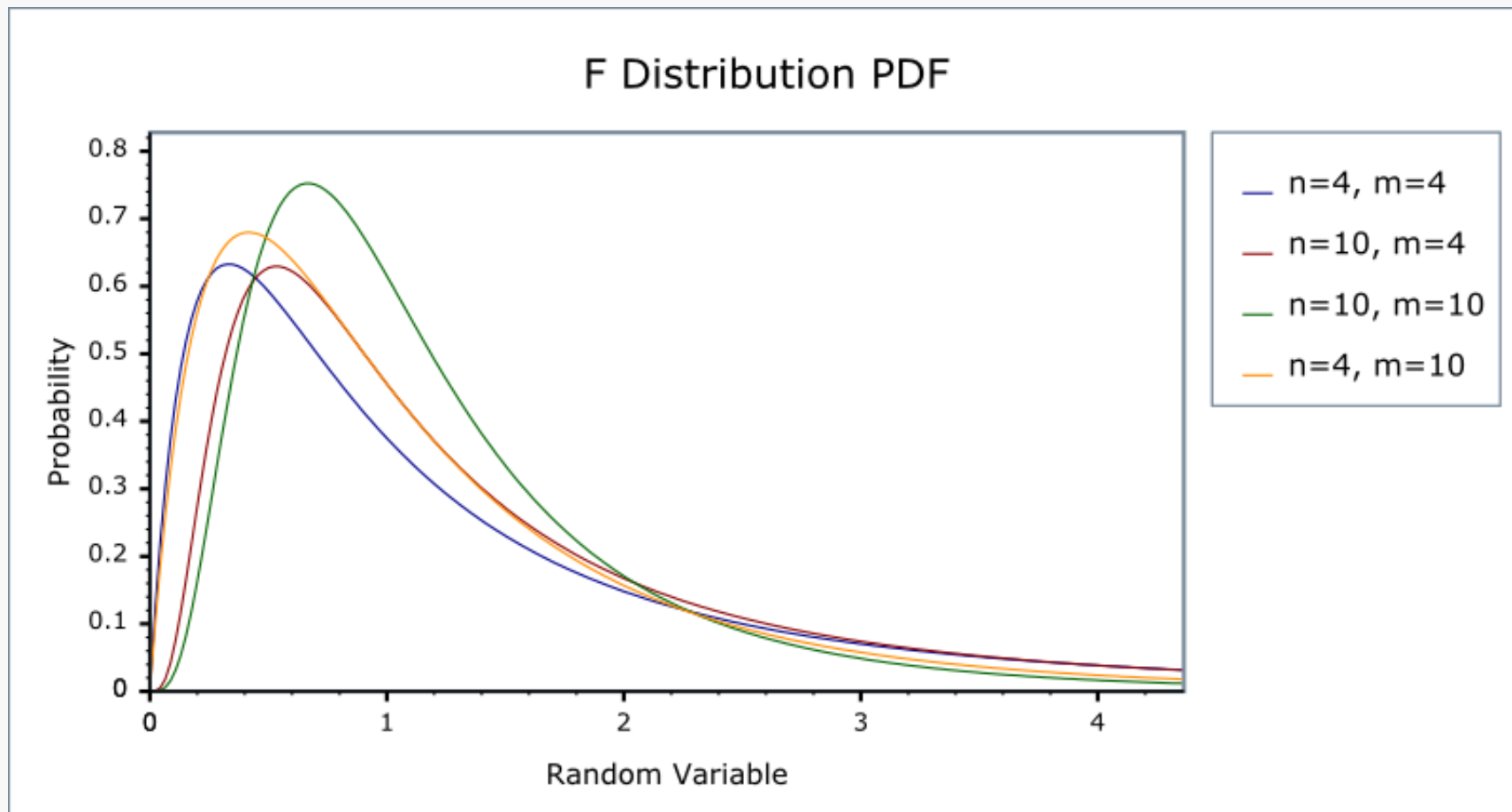
二、两个独立样本总体均值方差比的区间估计



自由度为 $n_1 - 1$ 、分母自由度为 $n_2 - 1$ 的 F 分布



二、两个独立样本总体均值方差比的区间估计



二、两个独立样本总体均值方差比的区间估计

由
$$P\left(\chi_{1-\alpha/2}^2 \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{\alpha/2}^2\right) = 1 - \alpha$$


$$P\left(\frac{s_1^2 / s_2^2}{F_{\alpha/2}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2 / s_2^2}{F_{1-\alpha/2}}\right) = 1 - \alpha$$

于是，可得总体方差比 σ_1^2 / σ_2^2 的 $1 - \alpha$ 估计区间为：

$$\left(\frac{s_1^2 / s_2^2}{F_{\alpha/2}}, \frac{s_1^2 / s_2^2}{F_{1-\alpha/2}} \right)$$



二、两个独立样本总体均值差比的区间估计

【例5.11】为研究男女青少年在体重上是否存在差异，某学校从13-15岁的学生中随机抽取25名男生和25名女生，调查并计算得到男生和女生样本体重的均值和方差分别为：男生： $\bar{x}_1 = 48.74$ ， $s_1^2 = 47.77$ ；女生： $\bar{x}_2 = 43.05$ ， $s_2^2 = 20.93$ 。试以90%的置信度，估计该校男女青少年体重方差比的置信区间。



二、两个独立样本总体均值差比的区间估计

解：总体方差 σ_1^2/σ_2^2 的90%的估计区间下限为：

$$\frac{s_1^2 / s_2^2}{F_{\alpha/2}} = \frac{47.77/20.93}{F_{0.05}} = \frac{2.28}{1.98} = 1.15$$

总体方差 σ_1^2/σ_2^2 的90%的估计区间上限为：

$$\frac{s_1^2 / s_2^2}{F_{1-\alpha/2}} = \frac{47.77/20.93}{F_{0.95}} = \frac{2.28}{0.505} = 4.51$$

可以推断总体方差比 σ_1^2/σ_2^2 在1.15到4.51之间，做出这一推断的把握程度为90%。



第八节 总体比率的区间估计

- 一、单样本总体比率的区间估计
- 二、两个独立样本总体比率差的区间估计
- 三、估计总体比率时样本容量的确定



一、单样本总体比率的区间估计

比率问题实际上属于二分类的定类变量问题。

总体比率 $P = \frac{N_1}{N}$, N_1 是总体中具有某种属性的个体单位数, N 为总体单位总数。

样本比率 $p = \frac{n_1}{n}$, n_1 是样本中具有某种属性的个体单位数, n 为样本容量。

数理统计证明: 当样本容量充分大时, 样本比率 p 服从:

$$N\left(P, \frac{P(1-P)}{n}\right)$$

由此可得 z 统计量: $Z = \frac{p-P}{\sqrt{P(1-P)/n}}$



一、单样本总体比率的区间估计

则给定置信度 $1 - \alpha$ 时：

总体比率 P 的区间估计为：

$$p \pm Z_{\alpha/2} \sqrt{\frac{P(1 - P)}{n}}$$

由于 P 未知，在大样本下用样本比率 p 来估计，即：

$$p \pm Z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}}$$

大样本的条件通常为： $np \geq 5$ 且 $n(1 - p) \geq 5$



一、单样本总体比率的区间估计

【例5.12】 为了解在校学生英语六级的考试通过率情况，某大学从该校某次参加六级考试的全部学生中随机抽取100名学生进行调查，其中有36名通过了英语六级考试。试以95%的置信度估计该校学生通过英语六级考试人数的比率。



一、单样本总体比率的区间估计

解： 样本比率为： $p = \frac{36}{100} = 0.36$

$$np = 100 \times 0.36 = 36 > 5 \text{ 且}$$

$$n(1-p) = 100 \times 0.64 = 64 > 5$$

为大样本，因此总体比率的95%估计区间为：

$$p \pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} = 0.36 \pm 1.96 \times \sqrt{\frac{0.36 \times (1-0.36)}{100}} = 0.36 \pm 0.094$$

即：[0.266, 0.454]，该校学生通过英语六级考试人数的比率在31.2%至40.8%之间，做出这一推断的把握程度为95%。



一、单样本总体比率的区间估计

基于SPSS的操作步骤:

- 打开数据，选择实现过程：
- Analyze→Compare Means→One-Sample T Test
- 选择变量：“exam” →Test Variable(s)
- 设置置信度：
Options→在“One-Sample T Test: Options”对话框中设定
95%→Continue→OK



一、单样本总体比率的区间估计

输出结果：

One-Sample Test			通过六级考试
Test Value = 0	t		7.462
	df		99
	Sig. (2-tailed)		.000
	Mean Difference		.360
	95% Confidence Interval of the Difference	Lower	.26
		Upper	.46

结果显示：在95%置信度下，该校学生通过英语六级考试的总体比率的估计区间为[0.26，0.46]。



二、两个独立样本总体比率差的区间估计

两个独立的随机样本，样本容量充分大时， $P_1 - P_2$ 服从：

$$N(P_1 - P_2, \sqrt{\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}})$$

由此可得Z统计量

$$Z = [(p_1 - p_2) - (P_1 - P_2)] / \sqrt{\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}}$$



二、两个独立样本总体比率差的区间估计

给定置信度时， $P_1 - P_2$ 的区间估计为：

$$(p_1 - p_2) \pm Z_{\alpha/2} \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

大样本情况下，以样本比率 p_1 替代总体比率 P_1 ，以样本比率 p_2 替代总体比率 P_2 ，则有：

$$(p_1 - p_2) \pm Z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$



二、两个独立样本总体比率差的区间估计

【例5.13】从两批产品中各随机抽取100件进行检查，结果第一批产品样本的合格率为95%，第二批产品样本的合格率为90%。试以95%的置信度估计两批产品合格率之差。



二、两个独立样本总体比率差的区间估计

解：根据公式有：

$$\begin{aligned}& (p_1 - p_2) \pm Z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\&= (0.95 - 0.90) \pm 1.96 \times \sqrt{\frac{0.95 \times (1-0.95)}{100} + \frac{0.90 \times (1-0.90)}{100}} \\&= 0.05 \pm 1.96 \times 0.037 \\&= 0.05 \pm 0.073\end{aligned}$$

即：[-0.023, 0.123]，两批产品合格率之差在-2.3%至12.3%之间，做出这一推断的把握程度为95%。



二、两个独立样本总体比率差的区间估计

基于SPSS的操作步骤：

- 打开数据，选择实现过程：
- Analyze→Compare Means→Independent-Samples T Test
- 选择变量：“qualified” →Test Variable(s)
- 设置置信度：

Options→在“Independent-Samples T Test: Options”对话框中
设定95%→Continue→OK



二、两个独立样本总体比率差的区间估计

输出结果：

Independent Samples Test			
		是否合格	
		Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F	7.445	
	Sig.	.007	
t-test for Equality of Means	t	1.342	1.342
	df	198	180.733
	Sig. (2-tailed)	.181	.181
	Mean Difference	.050	.050
	Std. Error Difference	.037	.037
	95% Confidence Interval of the Difference	Lower	-.023
		Upper	.124

结果显示：在95%置信度下，两种产品合格比率差的估计区间为[-0.024， 0.124]。



三、估计总体比率时样本容量的确定

给定极限误差 Δ 时，

$$\text{单个总体: } n = \frac{(Z_{\alpha/2})^2 P(1-P)}{\Delta^2}$$

$$\text{两个独立总体: } n_1 = n_2 = \frac{(Z_{\alpha/2})^2 P_1(1-P_1) + P_2(1-P_2)}{\Delta^2}。$$

实际中 Δ 多数时取0.10。

P 通常未知。一般按以下方法确定其估计值：

- ①用样本比率估计；
- ②抽取一个初始样本，调查初始样本的 S_P^2 ；
- ③遵循宁多勿少，取 $P = 0.5$ ，此时 $P(1 - P)$ 最大为0.25。



【例5.14】 根据以往的统计数据，已知某种产品的合格率为95%，现要求在极限误差不超过5%、置信度不小于95%的条件下，估计该种产品的合格率，至少应当抽取多大容量的样本？如果生产该产品共有两条生产线，欲以95%的置信度估计两条生产线上的产品合格率之差，那么至少应各抽取多大容量的两个独立样本？



解：已知 $\Delta_p = 5\%$, $Z_{\alpha/2} = 1.96$, $\sigma_p^2 = P(1-P) = 0.0475$

$$n = \frac{Z_{\alpha/2}^2 P(1-P)}{\Delta_p^2} = \frac{1.96^2 \times 0.0475}{(0.05)^2} = 72.99 \approx 73$$

即，为估计该种产品的合格率至少应抽取73件。

$$\begin{aligned} n_1 = n_2 &= \frac{\left(Z_{\alpha/2}\right)^2 [P_1(1-P_1) + P_2(1-P_2)]}{\Delta^2} \\ &= \frac{1.96^2 \times [0.5 \times (1-0.5) + 0.5 \times (1-0.5)]}{0.1^2} = 192.08 \approx 193 \end{aligned}$$

即，为估计两条生产线上的产品合格率之差，至少应从两条生产线上各抽取193件产品。



第九节 抽样组织方式

- 一、简单随机抽样
- 二、分层抽样
- 三、等距抽样
- 四、整群抽样
- 五、多阶段抽样



抽样估计

按照**随机原则**从调查对象中抽取一部分单位进行调查，并以调查结果对总体数量特征作出具有一定可靠程度的估计与推断，从而认识总体的一种统计方法。

指样本单位的抽取不受主观因素及其他系统性因素的影响，每个总体单位都有均等的被抽中机会



抽样估计的特点：

- 按随机原则抽取样本单位
- 目的是推断总体的数量特征
- 抽样推断的结果具有一定的可靠程度，抽样误差可以事先计算并控制

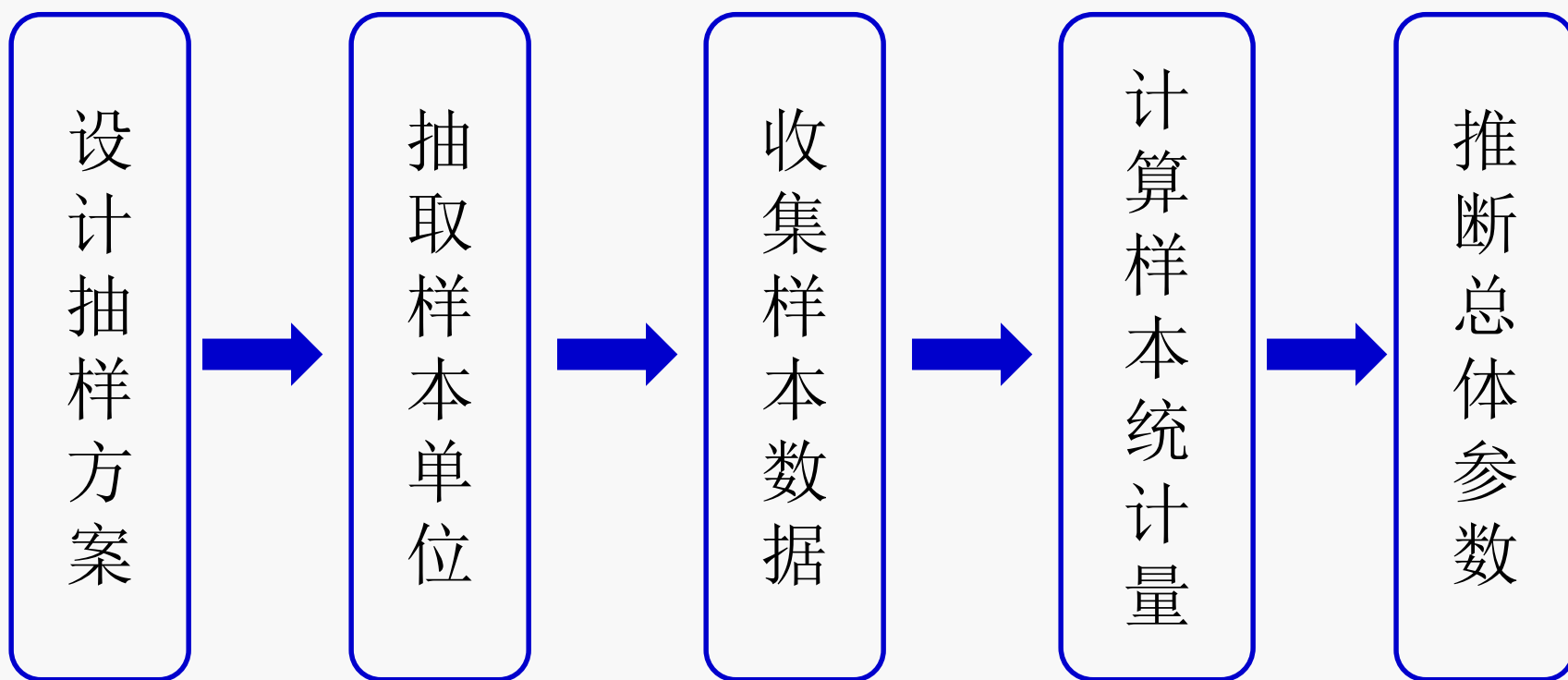
抽样估计的应用：

- 不可能进行全面调查时
- 不必要进行全面调查时
- 来不及进行全面调查时
- 对全面调查资料进行补充修正时



一、抽样分布的含义

抽样估计的一般步骤：



一、简单随机抽样

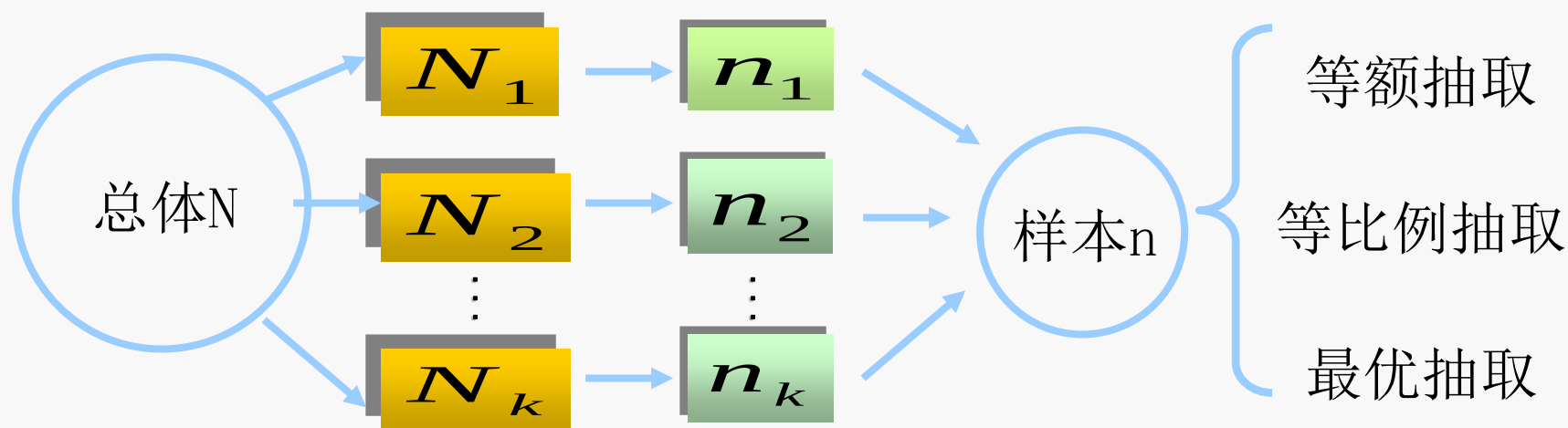
简单随机抽样 也称为纯随机抽样，对总体单位逐一编号，然后按随机原则直接从总体中抽出若干单位构成样本，保证总体中每个单位的中选机会相等。

- 仅适用于规模不大、内部各单位标志值差异较小的总体
- 是最简单、最基本、最符合随机原则，但同时也是抽样误差最大的抽样组织形式



二、分层抽样

分层抽样 也称为类型抽样，将总体全部单位分类，形成若干个类型组，然后从各类型中按随机原则分别抽取样本单位组成样本



- 能使样本结构更接近于总体结构，提高样本的代表性
- 能同时推断总体指标和各子总体的指标



三、等距抽样

等距抽样 也称为机械抽样或系统抽样, 将总体单位按某一标志排序, 而后按一定的间隔抽取样本单位

● 随机起点 ▲ 半距起点 ▼ 对称起点



(总体单位按某一标志排序)

- 按无关标志排队, 其抽样效果相当于简单随机抽样
- 按有关标志排队, 其抽样效果相当于分层抽样
- 等距抽样的误差一般较简单随机抽样的误差小

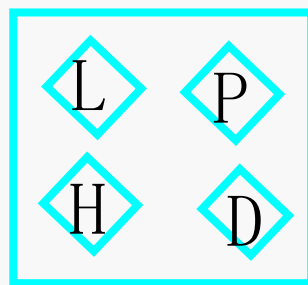
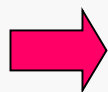
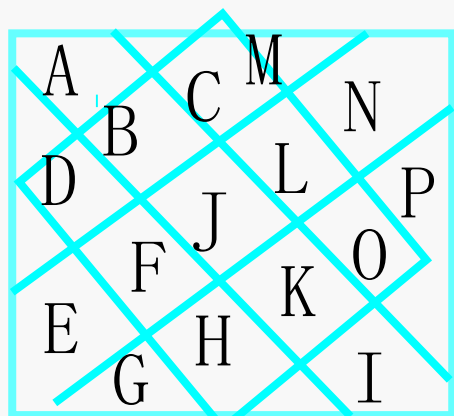


四、整群抽样

整群抽样 也称为聚类抽样，将总体全部单位分为若干个“群”，然后随机抽取一部分“群”，被抽中群体的所有单位构成样本

例：总体群数 $R=16$

样本群数 $r=4$



样本容量

$$n = n_d + n_p + n_l + n_h$$

- 简单、方便，能节省人力、物力、财力和时间，但其样本代表性可能较差。
- 如果总体单位多，分布区域广，缺少进行抽样的样本框，或按经济效益原则不易编制抽样框时宜采用。

五、多阶段抽样

多阶段抽样 又称多级抽样，是将整个抽样程序分解成若干阶段，然后逐阶段进行抽样，完成整个抽样过程

例：在某省100多万农户抽取1000户调查农户生产性投资情况：

第一阶段：从该省所有县中抽取5个县

第二阶段：从被抽中的5个县中各抽4个乡

第三阶段：从被抽中的20个乡中各抽5个村

第四阶段：从被抽中的100个村中各抽10户

样本 $n=100 \times 10=1000$ (户)



