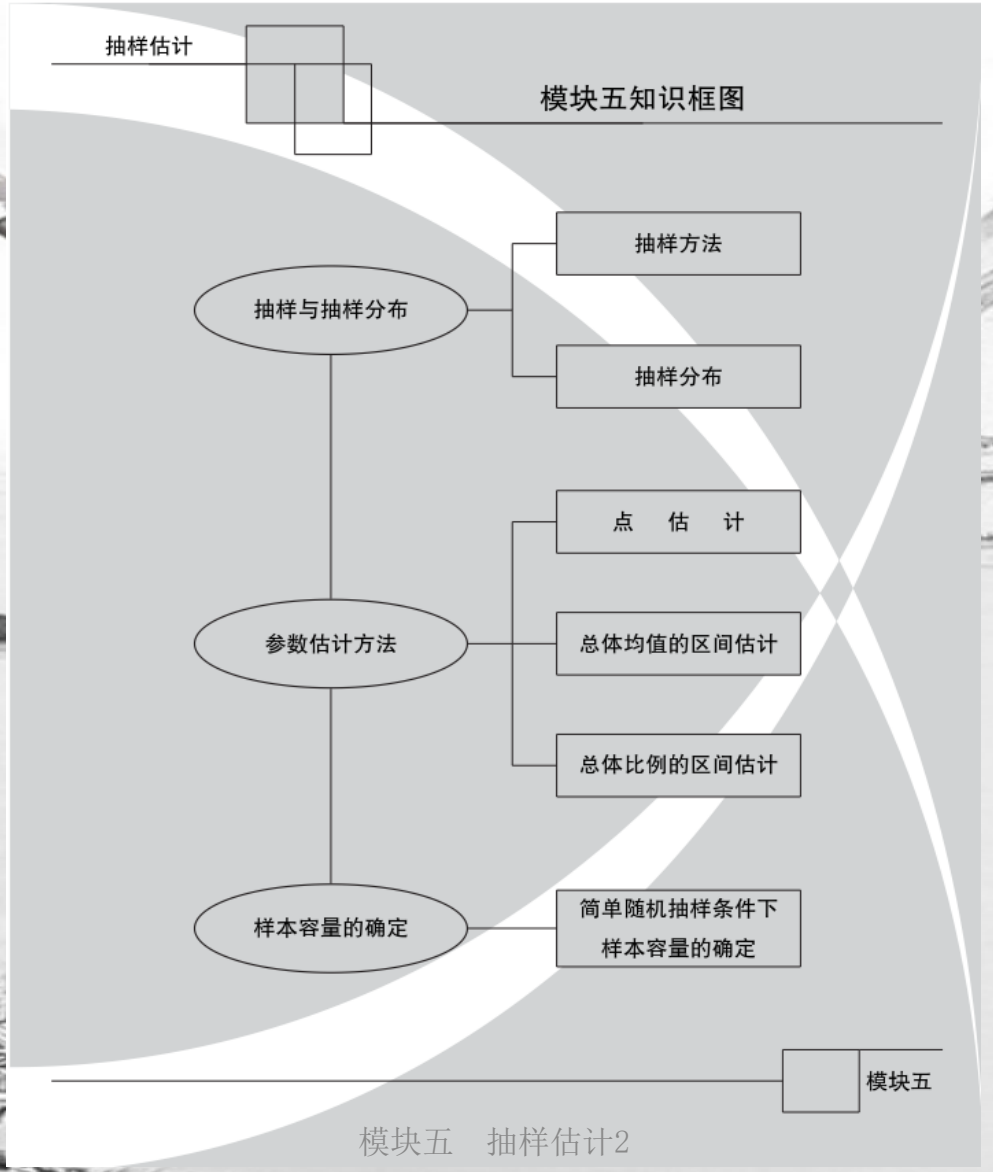
模块五　抽样估计

任务 1　抽样与抽样分布

任务 2　总体均值的区间估计

任务 3　总体比例的区间估计

任务 4　必要样本量的确定


抽样估计
模块五知识框图
抽样与抽样分布
抽样方法
抽样分布
参数估计方法
点　估　计
总体均值的区间估计
总体比例的区间估计
样本容量的确定
简单随机抽样条件下
样本容量的确定
模块五

知识目标

- 了解抽样方法的选择
- 了解抽样分布原理

能力目标

- 能够恰当地选择抽样方法并实施抽样

任务引入

某大学经管学院希望了解在校大学生的消费水平和消费结构。学院共有 2 200 名学生，要求随机抽取 40 名学生作为样本，应当怎样随机抽取这 40 名学生呢?

任务分析

在市场调查工作中，为了获得研究对象总体的数量特征值，可以采用普查的方法。但很多时候，不可能实施普查或普查在时间、人力、物力、财力上不够经济。这时，通常选择抽样估计的方法，即从总体中随机抽选一部分个体构成样本，计算样本的综合特征值，用样本信息去推算总体指标。本任务完成的是抽样估计的第一个环节———抽取样本。同时介绍抽样分布的有关知识，为下一个任务———参数估计的学习打下基础。

相关知识

一、抽样估计

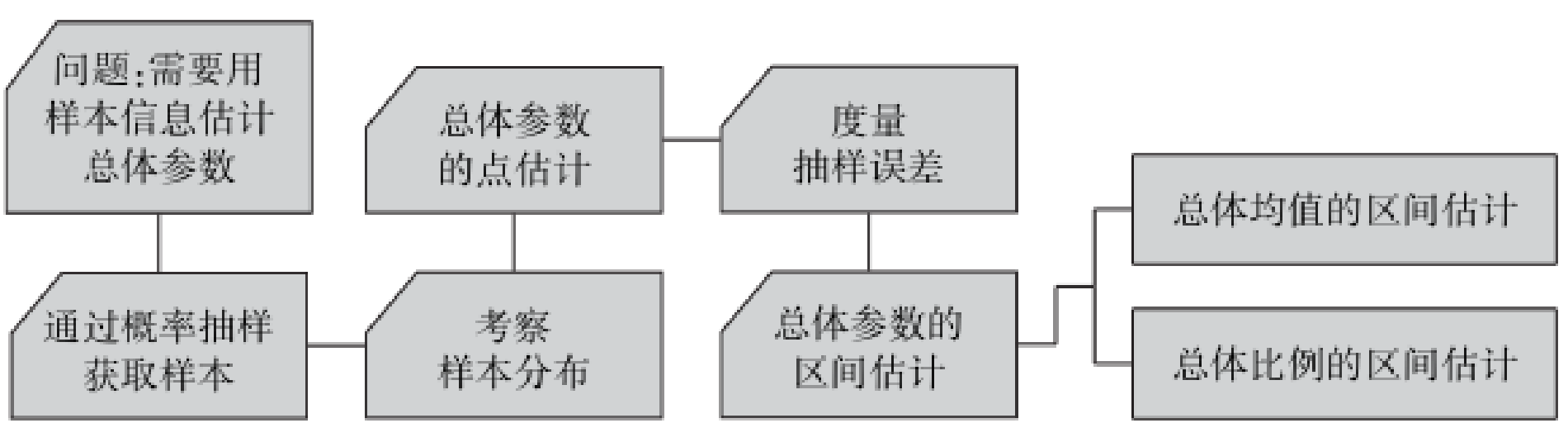
抽样估计是指在随机抽样的基础上，利用样本的实际资料计算样本统计量，并以样本统计量对总体参数作出具有一定可靠程度估计的一种统计分析方法。

抽样估计具有以下几个特点：

1. 是一种通过部分认识总体的统计分析方法。

2. 以概率抽样为基础，按随机原则抽取样本。

3. 可以用一定的概率将估计误差控制在一定的范围之内。

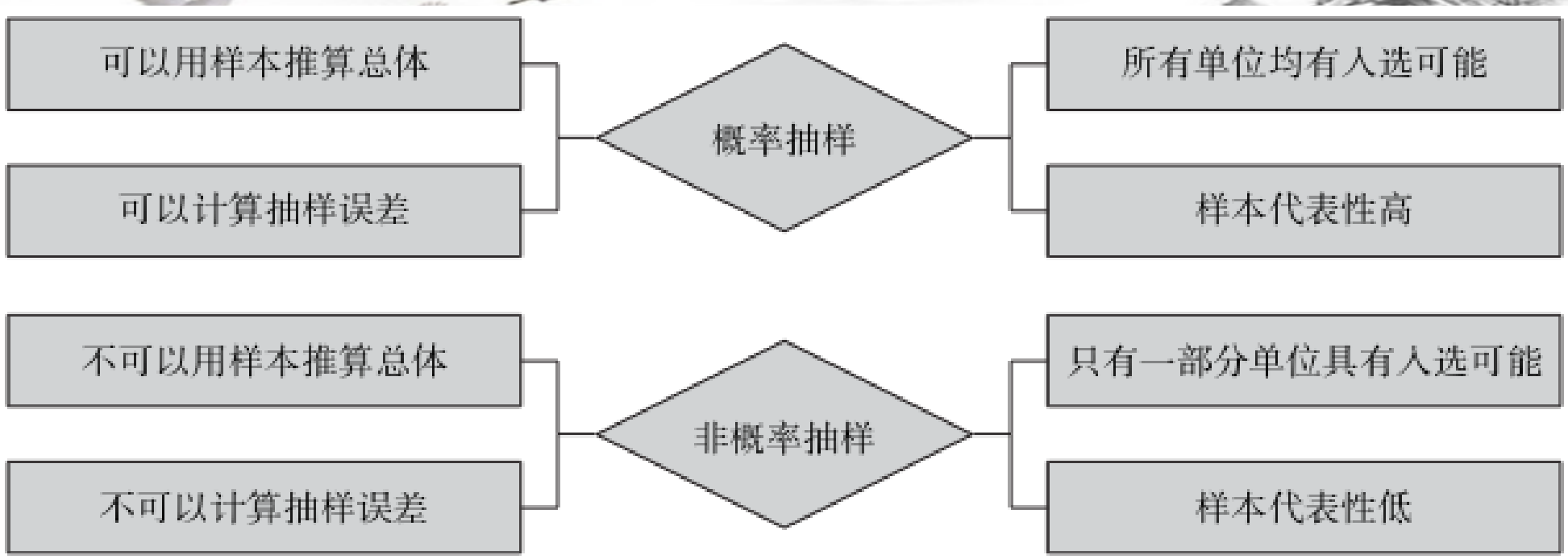

问题:需要用
样本信息估计
总体参数
总体参数
的点估计
度量
抽样误差
总体均值的区间估计
通过概率抽样
获取样本
考察
样本分布
总体参数的
区间估计
总体比例的区间估计

抽样估计要点图解

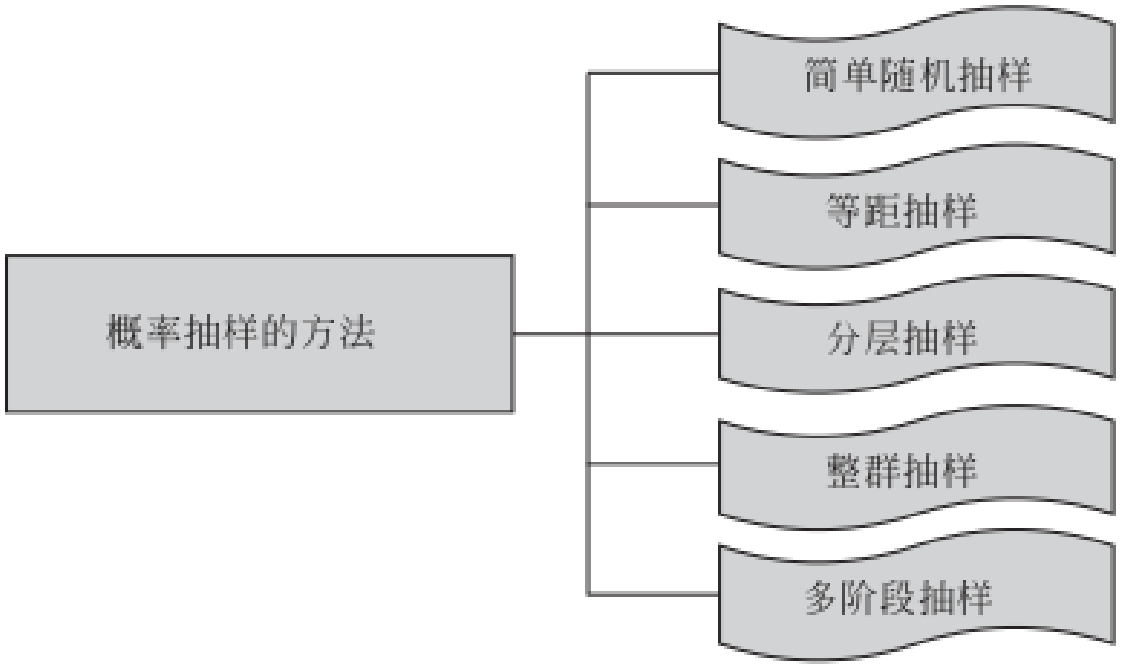
二、概率抽样方法

概率抽样又称为等概率抽样或随机抽样，是调查者按照随机原则抽取样本的方法。

非概率抽样又称为不等概率抽样或非随机抽样，是调查者根据自己的方便或主观判断抽取样本的方法。



概率抽样与非概率抽样的区别



抽样方法不同，样本统计量的计算方法也不同

概率抽样方法

1. 简单随机抽样

简单随机抽样是按随机原则直接从总体 N 个单位中抽取 n 个单位组成样本，总体中每个单位都有被抽中的机会。

简单随机抽样分两种。

(1) 重复抽样

重复抽样也称回置抽样，是指每次抽取一个样本单位登记后再放回总体中参加下一次抽取的方法，每一个样本单位都有被重复抽中的可能。

(2) 不重复抽样

不重复抽样也称不回置抽样，是指每次抽取一个样本登记后不放回总体中参加下一次抽取的方法，每一个样本单位只有一次被抽取的可能。

简单随机抽样的优点是当总体单位数不大或总体容量虽然较大但比较集中时，采用简单随机抽样容易取得较好的抽样效果。

2. 等距抽样

等距抽样又称系统抽样或机械抽样，是将总体各单位按一定标志或次序排列，然后按相等的距离或间隔抽取样本单位。

系统抽样两种抽取方式。

（1）等概率系统抽样

等概率系统抽样是指每个单位被抽中的概率是相等的。

（2）不等概率系统抽样（PPS 系统抽样）

不等概率系统抽样是指每个单位被抽中的概率是与该单位的规模成比例的。

3. 分层抽样

分层抽样也称类型抽样，先将总体所有单位按与研究内容密切相关的主要因素分成若干层，然后在各层中按随机原则抽取一定数量的单位构成样本。

分层抽样的常用方法有比例抽样法和加权比例抽样法两种。

（1）比例抽样法

比例抽样法是按照每层单位数在总体中所占的比例抽取样本单位数，适用于层与层之间变异程度大，各层内部变异程度不大的总体。各层的抽样单位数为：

$$n_i = \frac{N_i}{N} \cdot n \qquad (i = 1、2、\cdots、k)$$

式中，N 是总体单位总数，N_i 是每层的单位数，n 是应抽取的样本单位总数，n_i 是各层应抽取的样本单位数，k 是分层的层数，$\frac{N_i}{N}$ 是总体中各层单位数占总体单位总数的比重。

(2) 加权比例抽样法

加权比例抽样法是以每层的单位数与层内的标准差结合作为权数确定每层应抽取样本数的方法。

各层的抽样单位数为：

$$n_i = n \cdot \frac{W_i \cdot s_i}{\sum W \cdot s}$$

式中，n 是应抽取的样本单位总数，n_i 是各层应抽取的样本单位数，W_i 是各层单位数占总体单位数的比重，s_i 是各层内部的标准差 $\frac{W_i \cdot s_i}{\sum W \cdot s}$ 是同时考虑到各组的单位数比重和标准差后确定的各层的权数。

4. 整群抽样

整群抽样是先将所有总体单位分割为若干小群组，然后从中随机抽取一部分群，对中选群中的所有单位实施全面调查的一种抽样方法。

优点是以群为单位抽取，简化了抽样的工作量，节省了调查费用，也方便了调查的实施。缺点是样本单位在总体中分布不均匀，因此抽样误差常常大于简单随机抽样。

5. 多阶段抽样

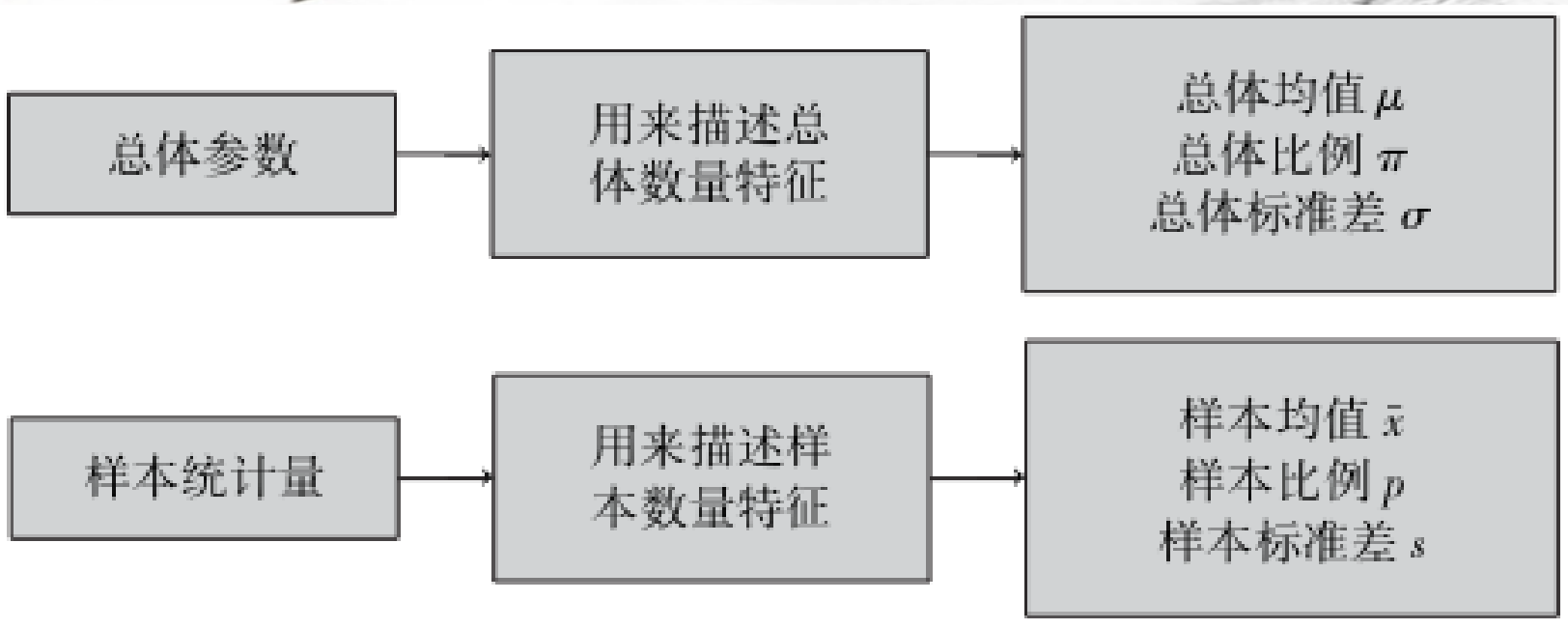
多阶段抽样又称为多级抽样，是指在抽取样本时，分为两个及两个以上的阶段从总体中抽取样本的一种抽样方式。

三、样本统计量的抽样分布

1. 几个基本概念

(1) 参数与统计量

总体参数是总体的综合特征值，总体参数通常是未知的，需要通过样本统计量推算获得。样本统计量是根据样本数据计算出的样本的综合特征值。常用总体参数与样本统计量的计算公式见表。


总体参数
用来描述总体数量特征
总体均值 μ
总体比例 π
总体标准差 σ
样本统计量
用来描述样本数量特征
样本均值 $\bar{x}$
样本比例 p
样本标准差 s

参数与统计量

总体参数与样本统计量的计算公式

		样本统计量	总体参数
均值	根据未分组资料计算	$\bar{x}=\frac{\sum x}{n}$	$\mu=\frac{\sum X}{N}$
	根据分组资料计算	$\bar{x}=\frac{\sum x\cdot f}{\sum f}$	$\mu=\frac{\sum X\cdot F}{\sum F}$
均值的标准差	根据未分组资料计算	$s_x=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$	$\sigma_x=\sqrt{\frac{\sum(X-\mu)^2}{N}}$
	根据分组资料计算	$s_x=\sqrt{\frac{\sum(x-\bar{x})^2\cdot f}{\sum f-1}}$	$\sigma_x=\sqrt{\frac{\sum(X-\mu)^2\cdot F}{\sum F}}$
比例		$p=\frac{n_0}{n}$；$1-p=\frac{n_1}{n}$	$\pi=\frac{N_0}{N}$；$1-\pi=\frac{N_1}{N}$
比例的标准差		$s_p=\sqrt{p\cdot(1-p)}$	$\sigma_p=\sqrt{\pi\cdot(1-\pi)}$

注：表 5—1—3 中符号代表的意义详见模块四。

(2) 样本容量和样本个数

样本容量是指一个样本所包含的样本单位数，一般用 n 表示。

样本个数是指从总体中可能抽取的样本个数。

如果采用重复抽样的方法，从总体N个单位中，随机抽取 n 个单位构成一个样本，则共可抽取 N^n 个样本。如果采用不重复抽样的方法，共可抽取 $\frac{N!}{n!\ (N-n)!}$ 个样本。

2. 抽样分布

抽样分布是指从某一总体中随机抽取容量为 n 的样本时，所有可能样本的统计量的频率分布或概率分布。

由表可知样本均值抽样分布的两个性质：

第一，样本均值 $\bar{x}_i$ 对称地分布在总体均值 $\mu = 2.5$ 的周围，大于 2.5 的样本均值出现的概率与小于 2.5 的样本均值出现的概率相等。

第二，重复抽样样本均值的方差大于不重复抽样样本均值的方差。

重复抽样和不重复抽样条件下的样本及样本统计量

抽样方法	重复抽样	不重复抽样
样本个数 k	16 个	6 个
所有可能的样本	1，1　2，1　3，1　4，1 1，2　2，2　3，2　4，2 1，3　2，3　3，3　4，3 1，4　2，4　3，4　4，4	1，2　2，3　3，4 1，3　2，4 1，4
样本均值 $\bar{x}_i$	1.0　1.5　2.0　2.5 1.5　2.0　2.5　3.0 2.0　2.5　3.0　3.5 2.5　3.0　3.5　4.0	1.5　2.5　3.5 2.0　3.0 2.5

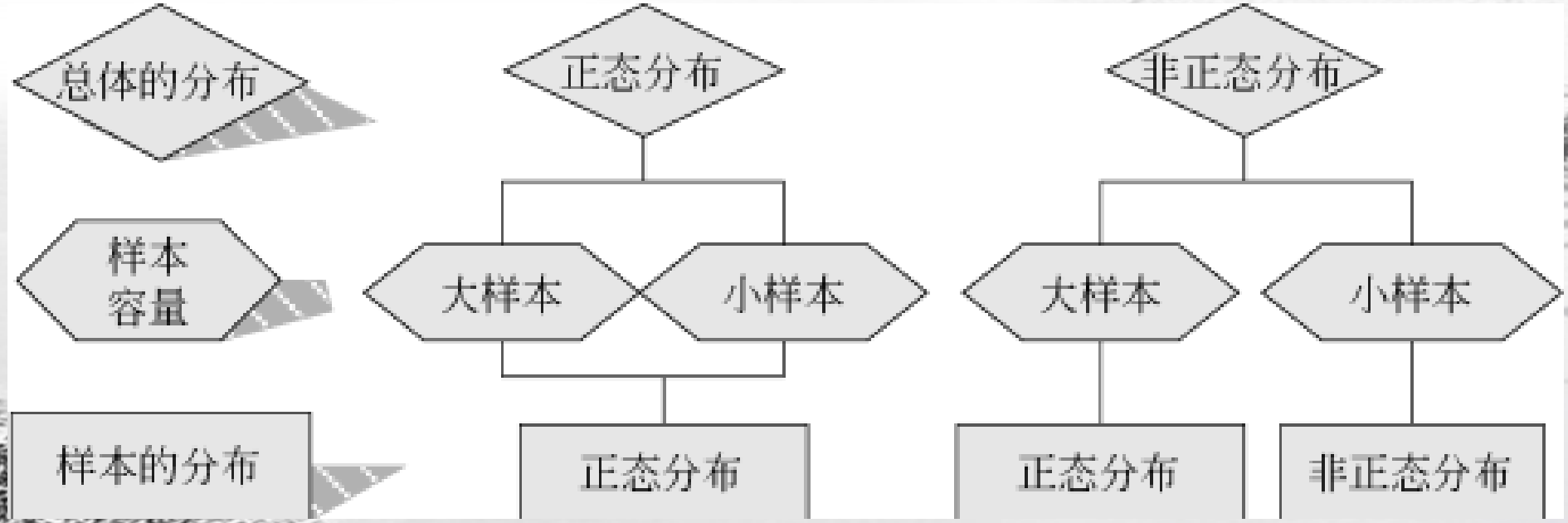
重复抽样和不重复抽样条件下的样本及样本统计量

抽样方法	重复抽样	不重复抽样
样本方差 s_i^2	0　0.5　2.0　4.5 0.5　0　0.5　2.0 2.0　0.5　0　0.5 4.5　2.0　0.5　0	0.5　0.5　0.5 2.0　2.0 4.5
所有样本均值的均值 $\mu_{\bar{x}}=\frac{\sum \bar{x}_i}{k}$	2.5	2.5
所有样本均值的方差 $\sigma_{\bar{x}}^2=\frac{\sum (\bar{x}_i-\mu)^2}{k}$	0.625	0.417

重复抽样和不重复抽样条件下样本均值的抽样分布

重复抽样			不重复抽样		
样本均值 $\bar{x}_i$	样本个数	发生的概率	样本均值 $\bar{x}_i$	样本个数	发生的概率
1.0	1	1/16	1.5	1	1/6
1.5	2	2/16	2.0	1	1/6
2.0	3	3/16	2.5	2	2/6
2.5	4	4/16	3.0	1	1/6
3.0	3	3/16	3.5	1	1/6
3.5	2	2/16			
4.0	1	1/16			
合计	16	1.0	合计	6	1.0

3. 均值的抽样分布与总体分布的关系



抽样分布与总体分布的关系

当总体服从正态分布时，无论样本容量大小，样本均值 $\bar{x}$ 均服从正态分布。样本均值的数学期望 $E(\bar{x})$ 等于总体均值 μ，样本均值的方差 $\sigma_{\bar{x}}^2$ 则与抽样方法有关。

重复抽样条件下：

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

不重复抽样条件下：

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

在大样本条件下，无论总体分布是否服从正态分布，样本均值的抽样分布均服从正态分布，样本均值 $\bar{x}$ 的数学期望 $E(\bar{x})=\mu$，样本均值 $\bar{x}$ 的方差 $\sigma_{\bar{x}}^2=\frac{1}{n}\sigma^2$ 。当 n 为小样本时 (通常认为 $n<30$ 为小样本)，样本均值的分布不服从正态分布，标准化的随机变量服从自由度为 (n - 1)的 t 分布。

在研究实际问题时，总体方差 σ^2 通常是未知的，可以用样本方差 s^2 代替总体方差 σ^2 (或用样本标准差 s 代替总体标准差 σ)，这样，样本均值的方差 $\sigma_{\bar{x}}^2=\frac{s^2}{n}$。

4. 样本比例的抽样分布

比例是指总体中具有某种属性或特征的单位数与总体单位数之比。若总体中具有某种属性的单位数为 N_1，不具有某种属性的单位数为 N_0，则将具有某种属性的单位数与全部单位数之比称为总体比例，即 $\pi=\frac{N_1}{N}$；样本中具有某种属性的单位数与全部单位数之比称为 p，$p=\frac{n_1}{n}$，$1-p=\frac{n_0}{n}$，$n=n_0+n_1$。

$1-\pi=\frac{N_0}{N}$，$N=N_0+N_1$

对于一个样本比例，如果 $n \cdot p \geq 5$ 和 $n \cdot (1-p) \geq 5$，就可以认为样本容量足够大。这时，样本比例 p 的期望值、抽样方差和抽样标准差为：

样本比例 p 的期望值：$E(p)=\pi$

样本比例的抽样方差 σ_p^2：

重复抽样条件下：

$$\sigma_p^2=\frac{\pi(1-\pi)}{n}$$

不重复抽样条件下：

$$\sigma_p^2=\frac{\pi(1-\pi)}{n}\cdot\frac{N-n}{N-1}$$

样本比例的抽样标准差 σ_p^2：

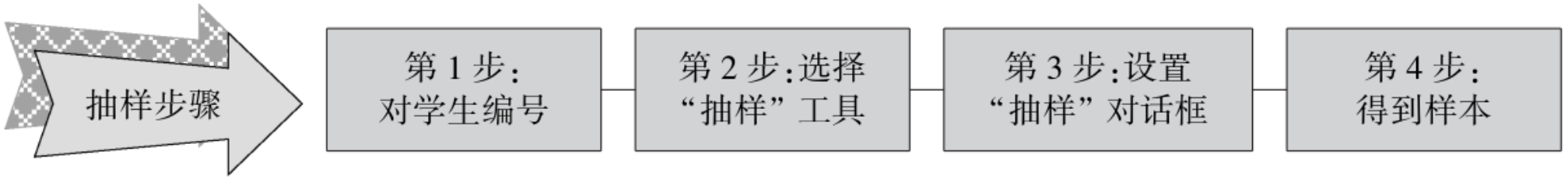
重复抽样条件下：

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

不重复抽样条件下：

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

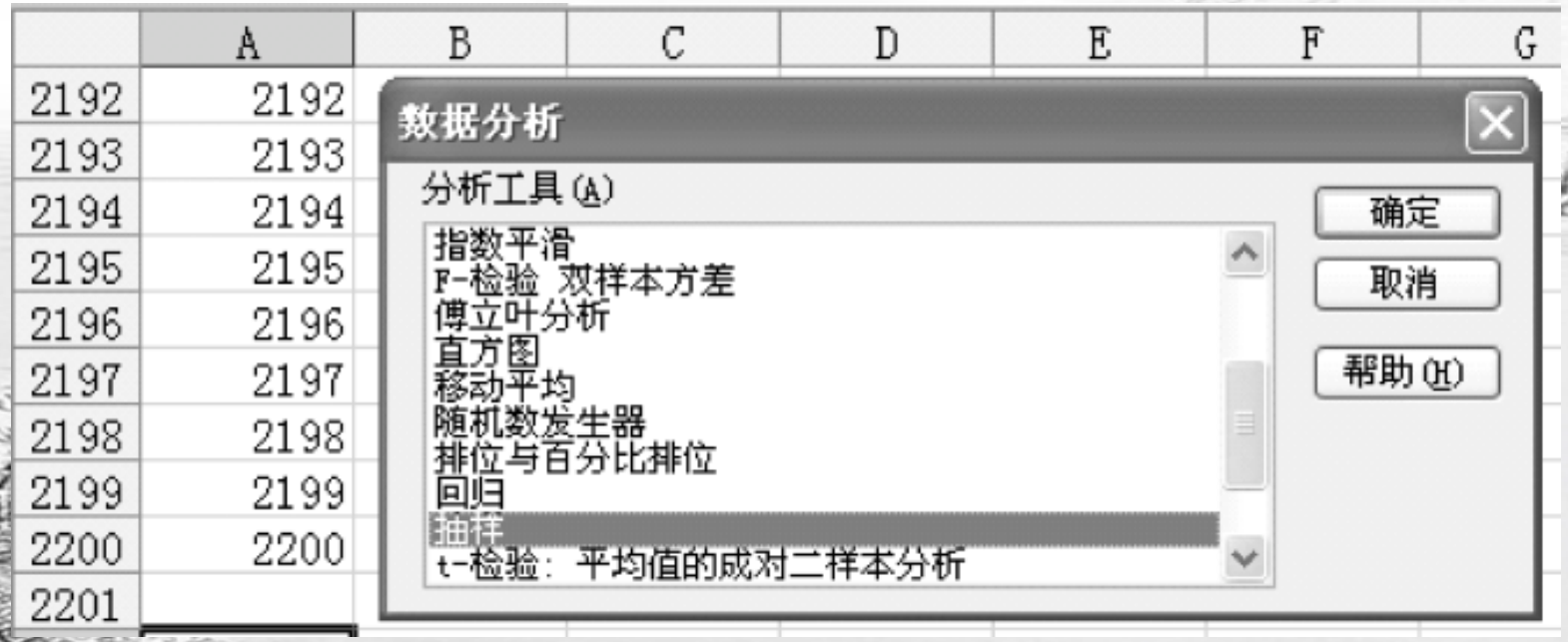
任务实施



利用 Excel 抽样的步骤

第 1 步：对该大学经管学院 2 200 名学生进行编号，从 1 号编至 2 200 号。

第 2 步：选择 “抽样” 工具：“工具” → “数据分析” → “抽样” → “确定”，如图所示。



Excel 中的 “抽样” 命令

第 3 步：设置 “抽样” 对话框并得到样本。

“抽样”对话框中，在“输入区域”输入学生编号所在单元格区域 “A1 ：A2200”；在“样本数”框中输入样本量“40”；在 “输出区域” 输入单元格 C1，如图所示。

设置 “抽样” 对话框

第 4 步：单击 “确定” 按钮，得到随机抽取的 40 名学生的编号，排序后如图所示。

	A	B	C	D
1	14	555	978	1544
2	74	561	1066	1568
3	82	602	1208	1586
4	149	624	1226	1830
5	207	675	1311	1848
6	214	753	1354	1935
7	305	800	1376	1946
8	369	808	1413	2086
9	549	809	1427	2101
10	550	835	1513	2110

随机抽出的 40 名学生的编号

知识目标

- 掌握参数估计的方法
- 掌握总体均值的区间估计

能力目标

- 能够熟练掌握区间估计的步骤
- 能够使用 Excel 函数进行区间估计

任务引入

模块五任务 1 中，利用 Excel 的随机抽样程序从 2 200 名学生中随机抽取了 40 名学生构成样本，现将这 40 名学生按每月手机话费金额排序得到表。要求根据所抽取学生的手机话费估计该大学经管学院 2 200 名学生的人均月手机话费，分别用 40名学生和其中 20 名学生的平均手机话费去估计学院全部学生的手机话费。

某大学经管学院 40 名学生每月手机话费金额 单位：元

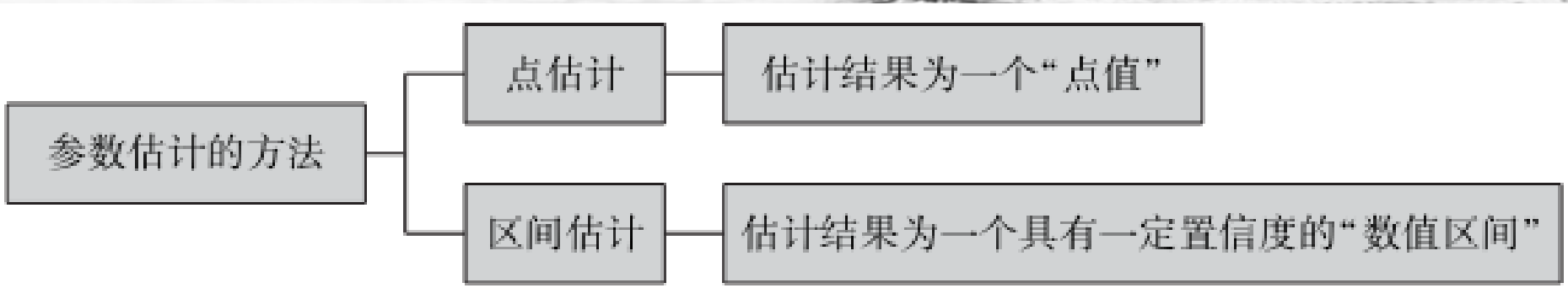
A	B	C	D	E	F
学生编号	手机话费	短信话费	学生编号	手机话费	短信话费
1	150	50	21	30	10
2	100	20	22	30	10
3	100	20	23	30	20
4	100	10	24	30	10
5	75	20	25	30	10
6	60	30	26	30	10
7	60	10	27	30	20
8	50	10	28	30	15
9	50	10	29	30	10
10	50	20	30	25	10
11	50	20	31	25	10
12	50	25	32	25	16
13	50	10	33	25	15
14	40	20	34	25	16
15	40	15	35	23	10
16	35	15	36	20	20
17	35	18	37	20	10
18	30	10	38	20	21
19	30	20	39	20	16
20	30	20	40	20	10

任务分析

样本抽取出来之后，就需要计算样本统计量并用样本统计量去估计总体参数。常用的样本统计量有样本均值、样本比例和样本方差，需要估计的总体参数相应的有总体均值、总体比例和总体方差，本任务的目的是引导大家学习怎样用样本均值去估计总体均值，以及怎样用样本均值去构造总体均值的置信区间。

相关知识

一、参数估计的方法



参数估计的方法

二、点估计

点估计是用某一个样本统计量的取值直接作为总体参数的估计值。

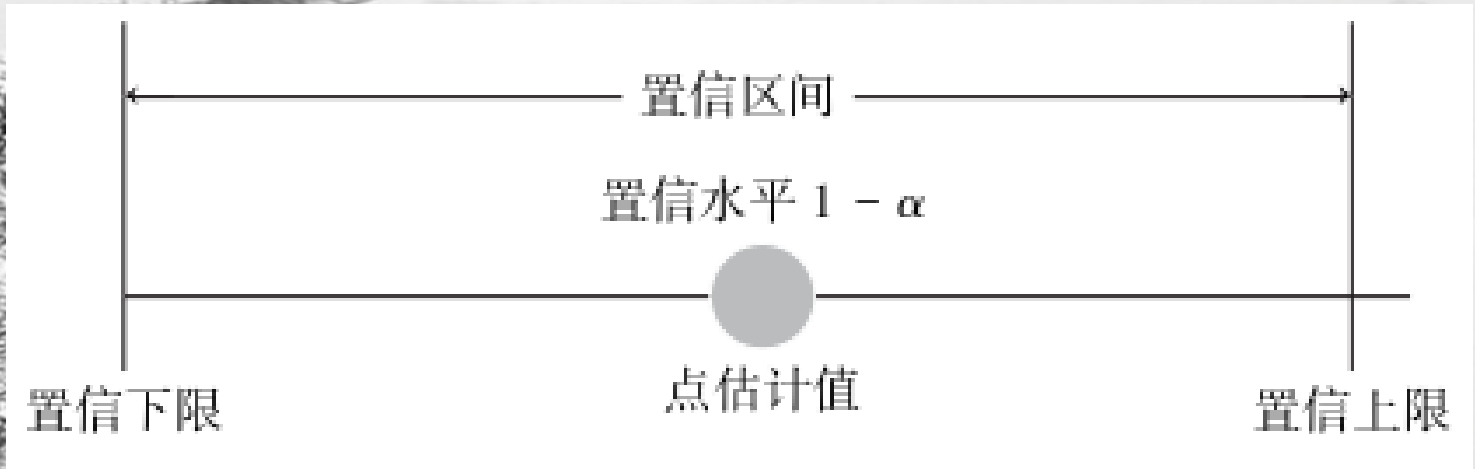
点估计的优点是简单明了，缺点是无法判断点估计的可靠性。但对于由点估计值构造的总体参数的置信区间，则可以给出估计的可靠程度。

三、总体均值的区间估计

1. 区间估计的基本原理

（1）区间估计

区间估计是在给定置信水平（1- α）的条件下，以点估计值为中心，构建总体参数的一个估计区间（或置信区间）。



置信区间

(2) 置信区间

置信区间是指在一定置信水平下总体参数的估计区间，其中，区间的最小值称为置信下限，最大值称为置信上限。置信区间可表示为：

点估计值±边际误差

边际误差也称为抽样极限误差或允许误差，是指在抽样估计时，根据所研究对象的变异程度和分析任务的要求确定的可允许的误差范围，它等于样本统计量可允许变动的上限或下限与总体参数之差的绝对值。边际误差的大小由两个因素决定：