第5章 数据收集

1 数据来源

2 传统数据收集方法

3 数据抽样及抽样估计

4 数据收集工具

5.1　数据来源

数据来源

- 数据的间接来源——二手数据
- 数据的直接来源——原始数据

数据的间接来源——二手数据的来源

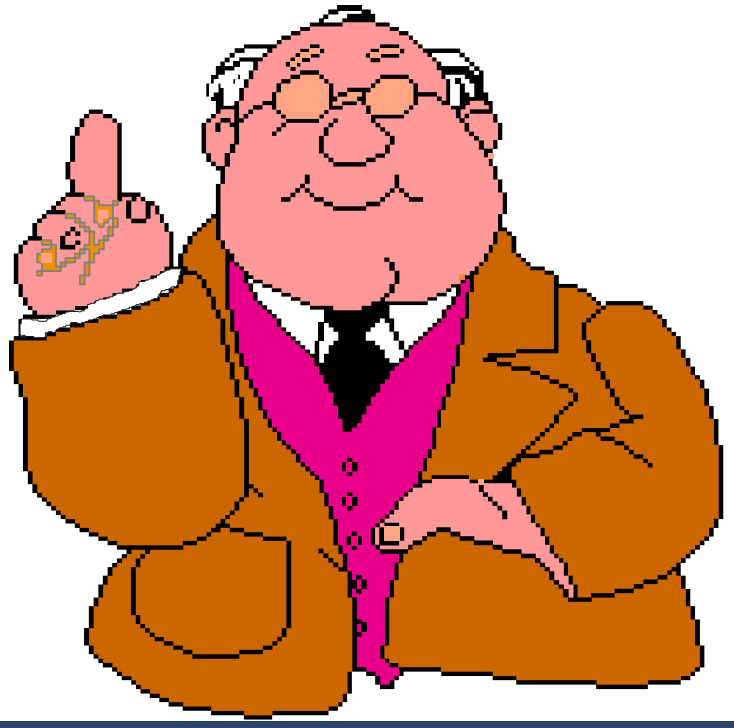
1. 统计部门和政府部门公布的有关资料，如各类统计年鉴
2. 各类经济信息中心、信息咨询机构、专业调查机构等提供的数据
3. 各类专业期刊、报纸、书籍所提供的资料
4. 各种会议，如博览会、展销会、交易会及专业性、学术性研讨会上交流的有关资料
5. 从互联网或图书馆查阅到的相关资料

二手数据的特点

1. 收集容易，采集成本低
2. 作用广泛
   - 分析所要研究的问题
   - 提供研究问题的背景
   - 帮助研究者更好地定义问题
   - 检验和回答某些疑问和假设
   - 寻找研究问题的思路和途径
3. 收集二手资料在研究中应优先考虑

二手数据的评估

1. 数据是谁收集的？
   - 可信度评估
2. 为什么目的而收集的？
3. 数据是怎样收集的？
4. 什么时候收集的？

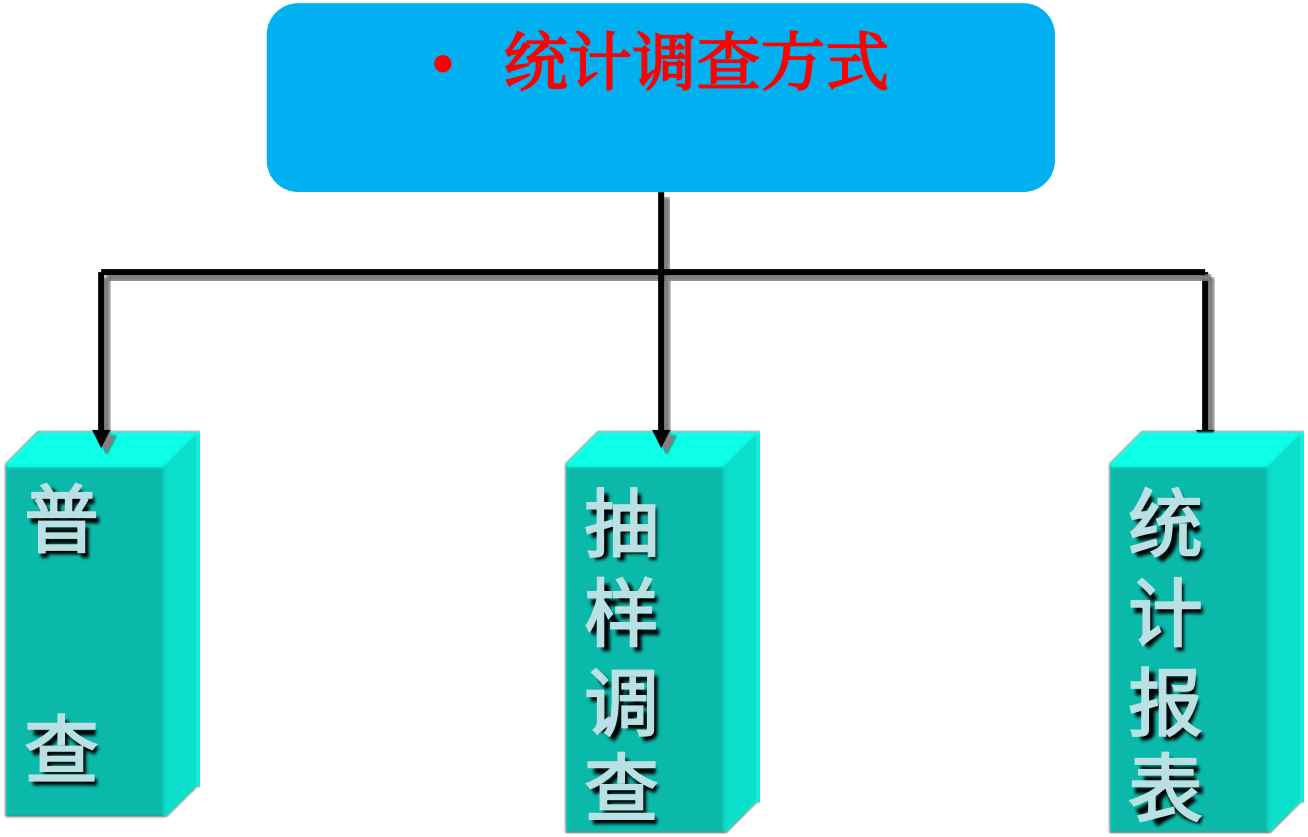


数据的直接来源
（原始数据）

1. 调查数据
   - 通过调查方法获得的数据
   - 通常是对社会现象而言
   - 通常取自有限总体
2. 实验数据
   - 通过实验方法得到的数据
   - 通常是对自然现象而言
   - 也被广泛运用到社会科学中
     - 如心理学、教育学、社会学、经济学、管理学等

5.2　传统数据收集方法

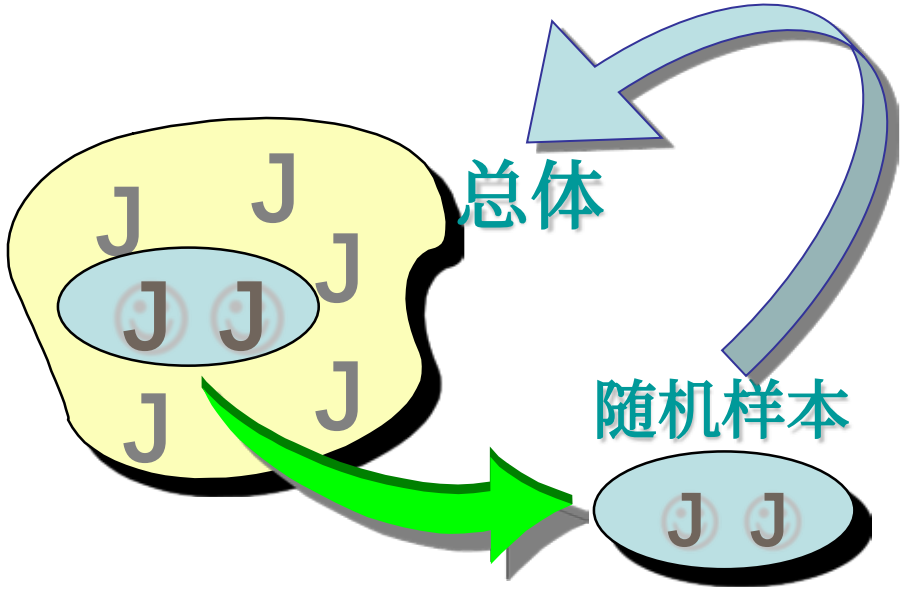
1.统计调查方式



2.抽样调查
(sampling survey)

1.从总体中随机抽取一部分单位作为样本进行调查，并根据样本调查结果来推断总体特征的数据收集方法

2．具有经济性、时效性强、适应面广、准确性高等特点



3.普查
(census)

1. 为特定目的专门组织的非经常性全面调查
2. 通常是一次性或周期性的
3. 一般需要规定统一的标准调查时间
4. 数据的规范化程度较高
5. 应用范围比较狭窄

4.统计报表
(statistical report forms)

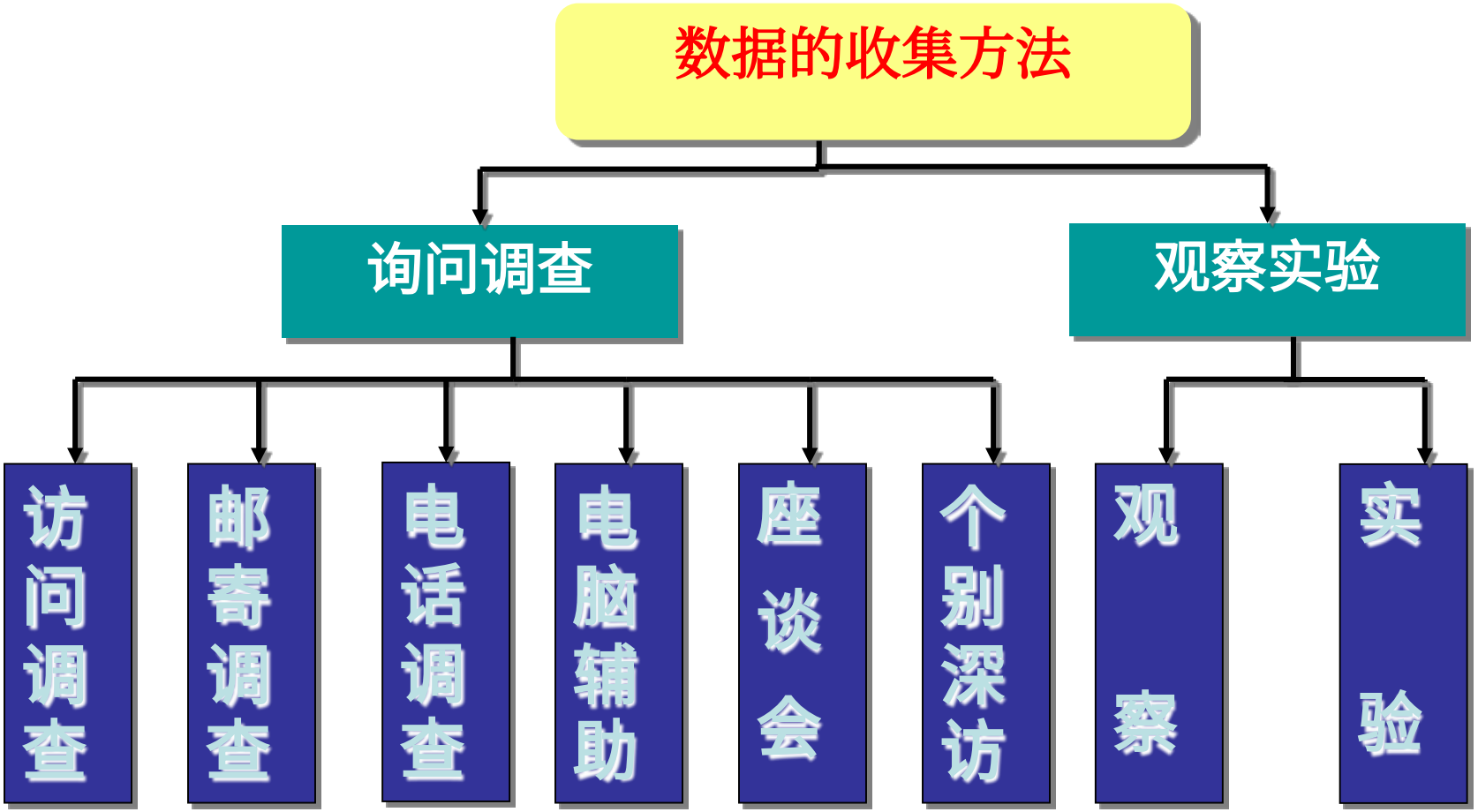
1.统计调查方式之一

2.过去曾经是我国主要的数据收集方式

3.按照国家有关法规的规定、自上而下地统一布置、自下而上地逐级提供基本统计数据

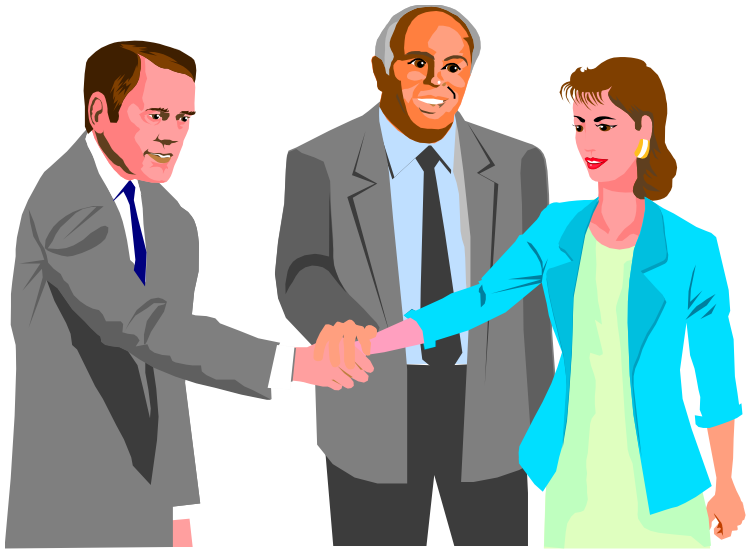
4.有各种各样的类型

5.数据的收集方法



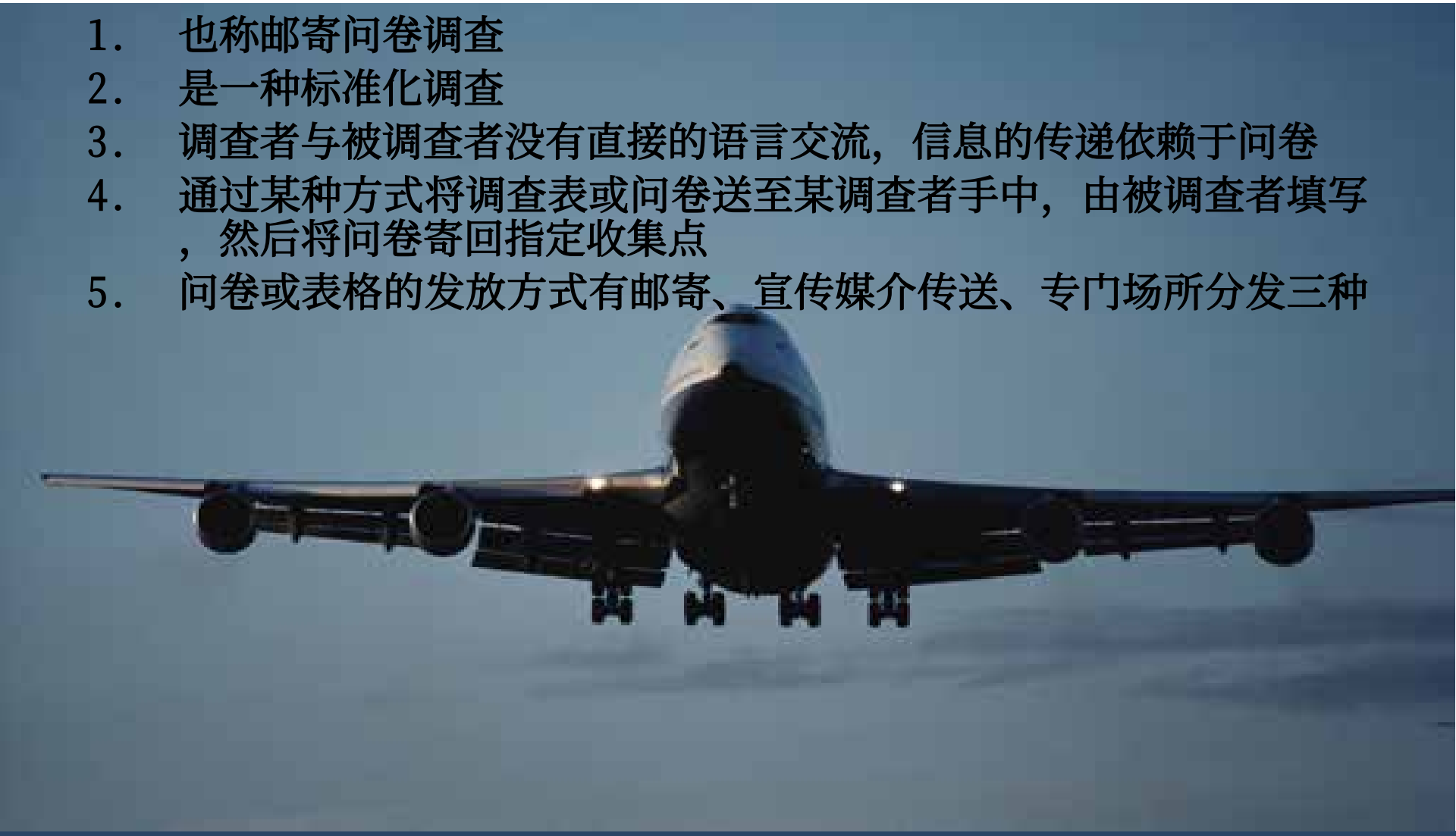
访问调查
(personal interview)

- 1．调查者与被调查者通过面对面地交谈而获得资料
- 2．有标准式访问和非标准式访问
  - 标准式访问通常按事先设计好的问卷进行
  - 非标准式访问事先一般不制作问卷



邮寄调查
(mail survey)

1. 也称邮寄问卷调查
2. 是一种标准化调查
3. 调查者与被调查者没有直接的语言交流，信息的传递依赖于问卷
4. 通过某种方式将调查表或问卷送至某调查者手中，由被调查者填写，然后将问卷寄回指定收集点
5. 问卷或表格的发放方式有邮寄、宣传媒介传送、专门场所分发三种



电话调查

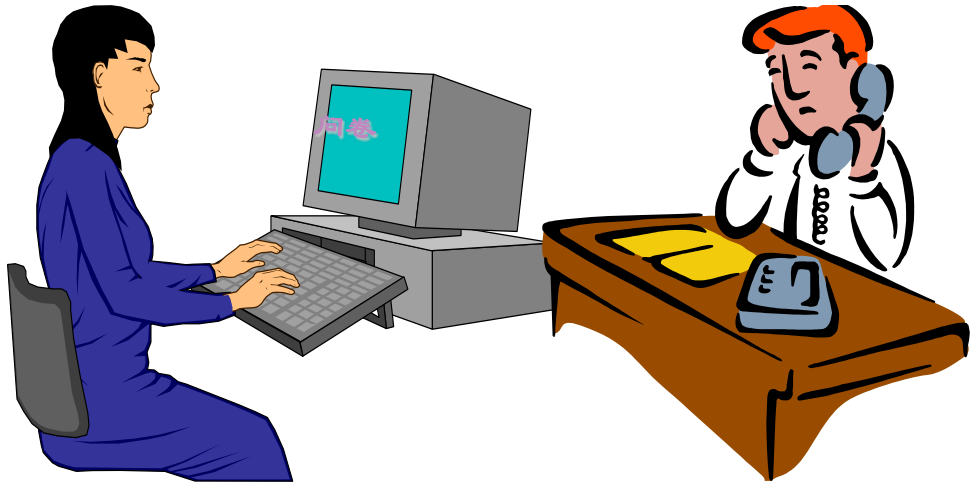
(telephone survey)

1. 调查者利用电话与被调查者进行语言交流以获得信息
2. 时效快、成本低
3. 问题的数量不宜过多

电脑辅助调查

(computer—assisted telephone interviewing)

1. 又称电脑辅助电话调查
2. 电脑与电话相结合完成调查的全过程
3. 一般需借助专门的软件进行
4. 硬件设备要求较高



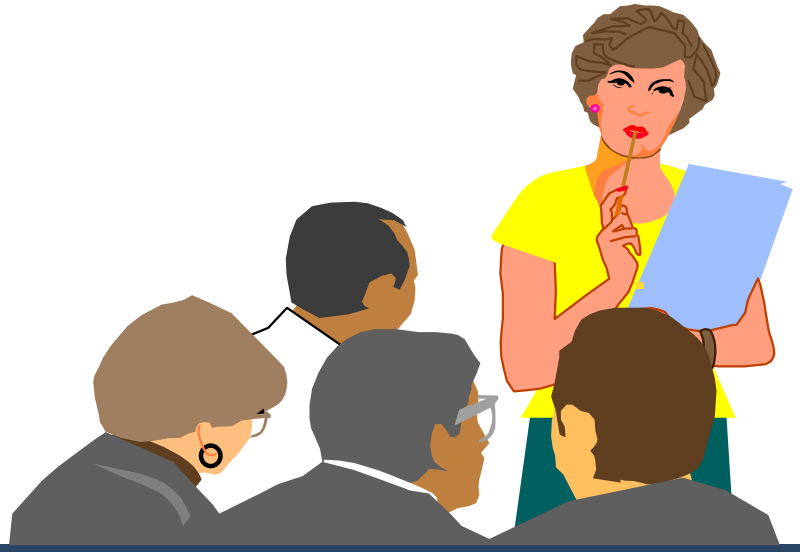
座谈会
(colloquia)

1.也称集体访谈

2.将一组被调查者集中在调查现场，让他们对调查的主题发表意见以获得资料

3.参加座谈会的人数不宜过多，一般为6～10人

4.侧重于定性研究



个别深度访问

(personal Interviewing)

1. 一次只有一名受访者参加、针对特殊问题的调查
2. 适合于较隐秘的问题，如个人隐私问题；或较敏感的问题，如政治方面的问题
3. 侧重于定性研究

观察法
(observational method)

- 1．就调查对象的行动和意识，调查人员边观察边记录以收集所需信息
- 2．调查人员不是强行介入
- 3．能够在被调查者不察觉的情况下获得资料

实验法
(experimental method)

- 1．在设定的特殊实验场所、特殊状态下，对调查对象进行实验以获得所需资料
- 2．有室内实验法和市场实验法

5.3　数据抽样及抽样估计

1.数据抽样及抽样估计

- 抽样估计的概念和特征

（一）抽样及抽样估计的概念

1.抽样即抽样调查，是指在总体中选取部分单位组成样本并收集样本单位的数据资料的过程。

2.抽样估计是在抽样调查的基础上，利用样本的数据资料计算样本指标，以样本特征值对总体特征值做出具有一定可靠程度的估计和判断。

(二)抽样估计的特点

1.抽样估计是由部分推断总体的一种认识方法

2.抽样估计建立在随机取样的基础上

3.抽样估计运用的是不确定的概率估计方法

4.抽样估计的误差可以事先计算并加以控制

二、抽样及抽样估计中的相关概念

（一）总体和样本

1.**总体：**是由**被调查对象的全部单位**所构成的集合体，简称总体。

总体容量：总体中的单位数，用N表示

2.**样本:**样本是从总体中**抽取的进行调查的部分单位**的集合体，又称抽样总体

样本容量：样本中的单位数，用n表示

大样本和小样本：n≥30时称大样本，n＜30称小样本

****应用:**在班级40名学生中随机选取15人进行健康状况调查,说明其中的总体、样本及容量

（二）概率抽样与非概率抽样

1.概率抽样:又称随机抽样，是按随机原则抽取样本单位。本章所指的均为概率抽样。

2.非概率抽样:又称非随机抽样，是指从研究的目的和需要出发，根据调查者的经验或判断，从总体中有意识地抽取部分单位构成样本。

**应用举例:重点调查、典型调查应为非概率抽样

（三）重复抽样和不重复抽样

1.重复抽样:又称**有放回的抽样**，从总体中抽取样本时，每次被抽中的单位都再被放回总体中参与下一次抽样。

2.不重复抽样:又称**无放回的抽样**，总体中随机抽选的单位经观察后不放回到总体中，即不再参加下次抽样。

（四）抽样框

1.概念:抽样框是包括全部抽样单位的名单框架。

2.形式

- 名单抽样框:如学生名单、职工名单、企业名单等
- 区域抽样框:如将一个城市按行政区划分为若干区、街道、居委会等
- 时间抽样框:如对流水线上的产品每隔一定时间抽取一定单位

（五）总体参数和样本统计量

1.总体参数:是**反映总体数量特征的数值**。在抽样推断 中，**参数是未知的、待估计的确定值**。

2.样本统计量:是根据样本资料计算的**反映样本数量特征的变量**，它的值随着样本的不同而变化，因此**是一个随机变量**。

表5-1　总体参数和样本统计量符号

总体指标符号	样本指标符号
总体容量： N	样本容量： n
总体平均数： m	样本平均数： $\bar{x}$
总体成数： P	样本成数： p
总体方差： s^2	样本方差： S^2
总体标准差： s	样本标准差： S

（六）抽样误差

1.统计误差及分类

统计误差

登记性误差: 统计调查中，由于观察、测量、登记、计算等原因或被调查者提供虚假信息所造成

代表性误差:以样本指标推断总体指标时产生的代表 性程度的差异。

偏差/系统误差：由于破坏随机原则而产生

随机性误差/抽样误差**：即使遵循随机原则以样本指标代表总体指标时的偏差

（六）抽样误差

2.抽样误差

抽样误差是指不包括登记性误差和系统性误差在内的随机误差，它衡量了抽样估计的精确度。

与抽样误差有关的三个概念

（1）**抽样实际误差**：指某一次具体抽样中，样本指标值与总体参数真实值之间的偏差。

（2）**抽样平均误差**：是指所有可能的样本指标与总体指标之间的平均差异程度，即**样本估计值的标准差**。

（3）**抽样极限/允许误差**：又称**置信区间**，是指一定概率下**抽样误差的可能范围**，说明样本估计量在总体参数周围变动的范围，记作Δ。

抽样平均误差

①抽样平均数的平均误差

概念：就是抽样平均数的标准差，反映抽样平均数的所有可能值对总体平均数的平均离散程度，记作 $s(\bar{x})$

定义公式：

$$s(\bar{x}) = \sqrt{\frac{S(\overline{x_i} - m)^2}{m}}$$

其中：$\overline{x_i}$ ：各个可能样本的平均数

m ：总体平均数

m ：重复抽样条件下所有可能的样本数

抽样平均误差

实际抽样推断中采用的公式

重复简单随机抽样 $$s(\bar{x}) = \sqrt{\frac{s^2}{n}}$$

不重复简单随机抽样 $$s(\bar{x}) = \sqrt{\frac{s^2}{n}(\frac{N-n}{N-1})}$$

其中，s^2 为总体方差

$\frac{N-n}{N-1}$ 为不重复抽样的修正因子

抽样平均误差

②样本成数（比例）的抽样平均误差

总体中具有某种特征的单位占全部总体单位数的比例称为总体比例，记作 P，样本中具有此种特征的单位占全部样本单位数的比例称为样本比例，记作 p 。

重复抽样条件下：

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

不重复抽样条件下：

$$s(p)=\sqrt{\frac{P(1-P)}{n}\left(1-\frac{n}{N}\right)}$$

抽样极限误差

①样本平均数的抽样极限误差：以绝对值形式表示的样本平均数的抽样误差的可能范围，用符号表示为：

$$|\bar{x} - m| \leq \Delta_{\bar{x}}$$

即：

$$m - \Delta_{\bar{x}} \leq \bar{x} \leq m + \Delta_{\bar{x}}$$

说明样本均值以确定的总体均值为中心，在 $m \pm \Delta_{\bar{x}}$ 之间变动。在实际抽样估计中是以样本均值推断总体均值的区间范围，因此，可将上述不等式做如下变换：

$$\bar{x} - \Delta_{\bar{x}} \leq m \leq \bar{x} + \Delta_{\bar{x}}$$

抽样极限/允许误差

② **样本比例的抽样极限误差：**以绝对值形式表示的样本比例的抽样误差的可能范围，用符号表示为：

$$|p - P| \leq \Delta_p$$

即：

$$P - \Delta_p \leq p \leq P + \Delta_p$$

同理，也可将上述不等式转换为：

$$p - \Delta_p \leq P \leq p + \Delta_p$$

2.抽样分布

一、抽样分布的概念和种类

（一）概念

抽样分布是样本统计量的概率分布。从一个总体中随机抽取容量相等的样本，根据样本资料计算某一统计量所有可能的概率分布，称为这个统计量的抽样分布。

（二）种类

精确分布／小样本分布：大多数是在正态分布总体条件下得到的，但应用不广

渐进分布／大样本分布：样本容量无限增大时统计量的极限分布，可看作是抽样分布的一种近似。

m

常见的抽样分布

（一）正态分布

1.正态分布：如果随机变量的概率密度函数为：

$$f(x)=\frac{1}{s\sqrt{2p}}e^{-\frac{(x-m)^2}{2s^2}} \qquad -\infty < x < +\infty$$

其中，m ,s 为常数且$s>0$，则称X服从参数为s、m的正态分布，记作 $X \sim N(m, s^2)$。

＊正态分布是最常见的抽样分布

常见的抽样分布

2. 标准正态分布：在正态分布中，当参数m=0，s =1时，则称X服从标准正态分布，记作 X~ $N(0$ ，1）。

标准正态分布的分布密度$j(x)$和分布函数$\Phi(x)$的性质如下：

（1）$j(x)$ 是偶函数，即 $j(-x)=j(x)$

（2）$\Phi(-x)$ =1-$\Phi(x)$

（3）如果 X ~ $N(m, s^2)$ ，则 X 的分布函数为

$$F(x)=\Phi\left(\frac{x-m}{s}\right)$$

上述公式称为正态分布函数的标准化公式。

常见的抽样分布

（二） c^2 分布

设 $x_1, x_2, \mathrm{L}, x_n$ 是独立同分布的随机变量，且每个随机变量都服从标准正态分布，即 $x_i \sim (0, 1)$，则随机变量 $c^2 = \sum_{i=1}^{n} x_i^2$ 的分布称为自由度为 n 的 c^2 分布，记作 $c^2(n)$。

当 $n \to \infty$ 时，c^2 分布趋近于正态分布，即 $c^2(n) \sim (n, 2n)$。

常见的抽样分布

（三）t分布

设随机变量 X与 Y相互独立，$X \sim (0, 1)$，$Y \sim c^2(n)$，则称随机变量 $t = \dfrac{X}{\sqrt{Y/n}}$

服从自由度为 n 的t 分布，记作$t(n)$。

当 $n \to \infty$时，t分布趋近于标准正态分布。实际应用中，当 $n > 30$时，t分布可用标准正态分布近似。

常见的抽样分布

（四）F 分布

1.设随机变量 X 与 Y 相互独立，且分别服从自由度为 n_1、n_2 的 c^2 分布，则称随机变量

$F=\frac{X/n_1}{Y/n_2}$ 服从第一自由度为n_1、第二自由度为 n_2 的 F 分布，记作 $F \sim F(n_1、n_2)$

2．F 分布对于两个总体的方差比的统计推断问题十分重要，F 是方差分析等统计推断方法的基础。与前两种分布不同的是 F 分布不以正态分布为其极限分布，它总**是一个正偏分布**。

3.抽样估计的基本方法

一、点估计

（一）概念

1.点估计

设总体随机变量的分布函数已知，但它的一个或多个参数未知，若从总体中抽取一组样本观察值，以该组数据来估计总体参数，就称为参数的点估计。

2.矩估计

矩估计法是用样本的矩去估计总体的矩，从而获得总体有关参数的估计量的方法。矩是指以期望值为基础定义的数字特征，如数学期望、方差、协方差等。

一、点估计

（二）矩估计法的评价

优点：1、 计算简便直观，一般不考虑抽样误差和可靠程度。

2、适用于对估计准确与可靠程度要求不高的情况

局限性： 1、它要求总体矩存在。

2、不能充分利用估计时已掌握的有关总体分布的信息。

（三）应用例题

[例7-1] 某厂对所生产的电子元件抽取5%进行抽样调查，计算出样本的平均耐用时间=4340小时，样本合格率=98%。根据矩估计法原理，估计该厂所生产的电子元件的平均耐用时间和合格率。

解：点估计法是用样本指标直接作为总体指标的代表值，所以，全部电子元件的平均耐用时间即为4340小时；总体合格率为98%。

二、区间估计

（一）区间估计的概念

根据样本统计量以一定的可靠程度去**估计总体参数值所在的范围**或区间，是抽样估计的主要方法。

（二）抽样估计的置信度与精确度

1.置信度：表示**区间估计的可靠程度或把握程度**，也即所估计的区间包含总体参数真实值的可能性大小，一般以1- α 表示。其中 α 表示显著性水平，即某一小概率事件发生的临界水平。

置信度通常采用三个标准：

（1）显著性水平=0.05，即1-α =0.95

（2）显著性水平=0.01，即1-α =0.99

（3）显著性水平=0.001，即1-α =0.999

（二）抽样估计的置信度与精确度**

2.**抽样估计的精确度**：用**置信区间的大小**即**抽样极限/允许误差**来表示

3.抽样估计的**置信度**与**精确度的矛盾关系**

在**样本容量和其他条件一定**的情况下，

若希望抽样估计有**较高的可靠度**，则必须扩大置信区间，即**必须降低估计的精确度**；

若希望抽样估计有**较高的精确度**，即置信区间范围缩小，则**必须降低估计的把握度**。

即：**抽样估计要求的把握度越高，则抽样允许误差越大，精确度越低；反之则相反**。

****思考**：在抽样调查中，如何同时提高抽样估计的精确度和把握度？

区间估计的应用

（一）总体均值的区间估计

1.总体方差已知时

当 $X \sim N(m, s^2)$时，来自该总体的简单随机样本 $x_1, x_2, \mathrm{L}, x_n$ 的样本均值服从数学期望为 m、方差 s^2为的正态分布，将样本均值统计量 $\bar{x}$

标准化，得到 Z 统计量 $Z = \dfrac{\bar{x} - m}{s / \sqrt{n}} \sim N(0,1)$

根据区间估计的定义，在给定的显著性水平a 下，总体均值 m 在 $1-a$ 的置信度下的置信区间为：

$\left(\bar{x} - Z_{a/2} \dfrac{s}{\sqrt{n}} , \bar{x} + Z_{a/2} \dfrac{s}{\sqrt{n}} \right)$，即 $\bar{x} - \mathrm{D}_{\bar{x}} \pounds m \pounds \bar{x} + \mathrm{D}_{\bar{x}}$

其中，$\dfrac{s}{\sqrt{n}} = s(\bar{x})$ 即抽样平均误差，$Z_{a/2} \dfrac{s}{\sqrt{n}} = \mathrm{D}_{\bar{x}}$ 即抽样允许误差

2.总体方差未知时总体均值的区间估计

**总体方差 s^2 未知，可以以样本方差 S^2 代替，但新的统计量不服从标准正态分布，而是服从自由度为 n -1的 t 分布

**给定置信度1- a，可查 t 分布表确定临界值 $t_{a/2}(n-1)$

从而总体均值的置信区间为：

$$\left(\bar{x} - t_{a/2}\frac{S_{n-1}}{\sqrt{n}} \quad , \bar{x} + t_{a/2}\frac{S_{n-1}}{\sqrt{n}}\right)$$

其中，$\frac{S_{n-1}}{\sqrt{n}} = s(x)$ 即为抽样平均误差

$t_{a/2}\frac{S_{n-1}}{\sqrt{n}} = \Delta_{\bar{x}}$ 即为抽样允许误差

上式也可表示为： $\bar{x} - \Delta_{\bar{x}} \le m \le \bar{x} + \Delta_{\bar{x}}$

（二）总体比例的区间估计

**在大样本下，样本比例的分布趋近于均值为 P 、方差为 $\frac{P(1-P)}{n}$

的正态分布。因此，给定置信度1- a ，查正态分布

表得 $Z_{a/2}$ ，则样本比例的抽样极限误差为：

$$D_p = Z_{a/2} \times s(p)$$

所以，总体比例的置信度为1- a 的置信区间为：

$$p - D_p \leq P \leq p + D_p$$

4.抽样调查的组织方式

一、简单随机抽样

(一)概念

又称纯随机抽样，是对总体单位不做任何分类或排队，直接从总体中按随机原则抽取样本单位的调查方式

(二)评价：简单易行，最符合随机原则，是抽样调查的基本形式

(三)适用情况：当总体单位数不多且分布比较均匀，或总体单位之间数量特征值差异较小，或总体单位有现成的编号时，采用这种方式比较适宜。

二、类型抽样

（一）概念

> 又称分层抽样或分类抽样，是将统计分组和抽样调查结合起来的组织方式。先将总体单位按某一标志分成若干组，然后在各组中采用简单随机抽样或其他方式抽取样本单位。

（二）适用情况：总体单位在被研究标志上有明显差异时。

（三）遵循原则：分组时应使组内差异尽可能小，组间差异尽可能大。

（四）种类：
- 等比例类型抽样
- 不等比例类型抽样

三、等距抽样

（一）概念

又称机械抽样或系统抽样，它是先将总体各单位按某一标志顺序排列，然后按照固定的顺序和相同的间隔抽取样本单位的抽样组织方式。

（二）分类

无关标志排序抽样：排序的标志与被研究的标志无关，实质是简单随机抽样。

有关标志排序抽样：排序的标志与被研究的标志有关，有利于提高样本的代表性 。

（三）评价：抽样误差一般较简单随机抽样小，当被研究现象标志变异程度较大时，更能显示出其优越性。但有可能产生系统性误差。

四、整群抽样

（一）概念

又称分群抽样或集团抽样，是将总体划分为若干群，然后以群为单位按简单随机抽样或等距抽样方式抽取部分群，对中选群中的所有单位一一调查的抽样组织方式。

（二）整群抽样与类型抽样的区别

类型抽样划分的组称为“类”，作用是缩小总体，使总体的变异减少，而抽取的基本单位仍是总体单位；

整群抽样划分的组称为“群”，作用是扩大单位，抽取的基本单位不是总体单位而是群，从而简化抽样工作程序。

（三）评价

样本单位集中于群内，显著地影响了总体单位分配的均匀性。与其他方式相比，在相同的条件下，**抽样误差较大，代表性较低。**

五、多阶段抽样

（一）概念

多阶段抽样又称为多级抽样，它是将抽取样本单位的过程划分为几个阶段，然后逐阶段抽取样本单位的抽样组织方式。

（二）优点

1.便于组织抽样

2.可以获得各阶段单元的调查资料

3.方式灵活

4.抽样单位的分布较广，降低抽样误差

（三）适用情况

当总体单位很多且分布广泛，几乎不可能从总体中直接抽取总体单位时，常采用多阶段抽样。

5.4　数据收集工具——IBM SPSS Data Collection

数据抽样工具——IBM SPSS Data Collection

1. IBM SPSS Data Collection 是一个数据收集工具。它以问卷为基础，支持多种方式包括 WEB、CAPI、CATI 来收集数据， 并且支持以多种数据格式存储来满足各种各样的用户需求。
2. 运用 IBM SPSS Data Collection 可以轻松地设计多样的调查问卷，精准地获取样本抽样，快速地收集有效数据并剔除不完整的数据回馈，方便地生成各种结果分析统计报告。
3. IBM SPSS Data Collection 支持三种问卷调查模式，基于 Web 的问卷收集，运用计算机辅助电话访谈的问卷收集以及离线的手持设备辅助的问卷调查。