

第二十五章抽样调查

1、抽样调查的几个基本概念（高频考点、特别关注）

抽样调查是指按照某种原则和程序，从总体中抽取部分单位，通过对这一部分单位进行调查得到信息，以达到对总体情况的了解，或者对总体的有关参数进行估计。

抽样调查是使用频率最高的一种调查方式。

以下示例分析研究某公司所有注册在职人员的工资状况，公司总人数为1000人，抽取100人调查。

（1）总体与样本（高频考点）

①总体——即调查对象的全体

如，公司所有注册在职员工。

②总体单元或单位——组成总体的各个个体

如，每个注册在职人员。

③样本——总体一部分

由总体中按一定原则或程序抽出的部分个体所组成。如，抽取的100个注册在职人员。

④入样单位——每个被抽中进入样本的单位如，抽取的每一个注册在职人员。

⑤样本量——样本中包含的入样单位的个数

如，抽取的100个注册在职人员，样本量是100个。

（2）总体参数与样本统计量

①总体参数——变量的数字特征

是根据总体中所有单位的数值计算的。

常用的总体参数有总体总量，总体均值，总体比例，总体方差等。

如，公司所有注册在职人员的平均工资。

②样本统计量（估计量）——对总体参数的估计

是根据样本中各单位的数值计算的，是对总体参数的估计，也称为估计量。样本统计量是一个随机变量，它取决于样本设计和正好被选入样本的单元特定组合。

常用的样本统计量有样本均值、样本比例、样本方差等。如，100名注册在职人员的平均工资。

（3）抽样框——供抽样所用的所有抽样单元的名单

是抽样总体的具体表现。抽样框可以有多种形式，常用的有名录框，也可以是一张地图或其他适当形式。如，公司1000名注册在职人员的人员名册。

但不管什么形式，抽样框中的单位必须是有序的，便于编号。高质量的抽样框应当提供被调查单位更多的信息，并且没有重复和遗漏。

2、概率抽样与非概率抽样（按抽取样本方法的不同）（高频考点）

|          | 定义                                | 特点或方法                                                                   |
|----------|-----------------------------------|-------------------------------------------------------------------------|
| 概率<br>抽样 | 也称随机抽样，是指依据随机原则，按照某种事先设计的程序，从总体中抽 | 特点：<br>①按一定的概率以随机原则抽取样本；②总体中每个单元被抽中的概率是已知的，或者是可以计算出来的；③当采用样本对总体参数进行估计时， |

|  |          |                   |
|--|----------|-------------------|
|  | 取部分单位的方法 | 要考虑到每个样本单位被抽中的概率。 |
|--|----------|-------------------|

|       | 定义                                                       | 特点或方法                                                                                                                                                                                                                                                                                     |
|-------|----------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 非概率抽样 | 又称为非随机抽样，是调查者根据自己的方便或主观判断抽取样本的方法，其最主要的特征是抽取样本时并不是依据随机原则。 | <p>主要方法：</p> <p>①<b>判断抽样</b>，指在抽取样本时，调查人员依据调查目的和对调查对象情况的了解，人为确定样本单元；</p> <p>②<b>方便抽样</b>，指在抽取样本时，依据方便原则，以达到最大限度降低调查成本的目的；</p> <p>③<b>自愿样本</b>，不是经过抽取，而是自愿接受调查的单元所组成的样本；网上调查比较典型。</p> <p>④<b>配额样本</b>，指将总体中的各单元按一定标准划分为若干类型，将样本数额分配到各类型中，从各类型中抽取样本的方法则没有严格限制，一般采用<b>方便抽样的方法</b>抽取样本单元。</p> |

### 3、抽样调查一般步骤

- (1) 确定调查问题——为什么要做、要做什么样的调查。
- (2) 调查方案设计——明确如何实施调查，抽样方案和问卷。
- (3) 实施调查过程。
- (4) 数据处理分析。
- (5) 撰写调查报告。

### 4、抽样调查中的误差（高频考点、特别关注）

|    |                        |                                                                                                     |
|----|------------------------|-----------------------------------------------------------------------------------------------------|
| 定义 | 样本估计值和总体参数真值之间的差异称为误差。 |                                                                                                     |
| 种类 | 抽样误差                   | 是由于 <b>抽样的随机性</b> 造成的，用样本统计量估计总体参数时出现的误差。抽到哪一个样本完全是随机的，而抽到不同的样本，对总体的估计就会不同，这就是抽样误差产生的 <b>根本原因</b> 。 |

|    |       |                                    |                                                                                                                                    |
|----|-------|------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|
| 种类 | 非抽样误差 | 是指除抽样误差以外，由其他原因引起的样本统计量与总体真值之间的差异。 |                                                                                                                                    |
|    |       | ① <b>抽样框误差</b>                     | 如把营业执照作为个体商业的抽样框，有些无照经营；有些虽有照但已转行；还有些一人多照。                                                                                         |
|    |       | ② <b>无回答误差</b>                     | 随机因素：被调查者不在家或有病无法接受调查。非随机因素：被调查者不愿告诉实情拒绝回答。                                                                                        |
|    |       | ③ <b>计量误差</b>                      | 由于调查所获得的数据与其真值之间不一致造成的误差。可能是由调查人员、问卷设计、受访者等原因造成的。<br>调查者诱导被调查者，调查中的提问错误或记录答案错误，调查人员有意作弊；由于问卷的原因受访者对调查问题的理解上有偏误；受访者记忆不清，受访者提供虚假数字等。 |

### 5、几种基本概率抽样方法

- (1) **简单随机抽样**；

- (2) 分层抽样;
- (3) 系统抽样;
- (4) 整群抽样;
- (5) 多阶段抽样。

## 6、简单随机抽样

(1) 简单随机抽样分为不放回简单随机抽样和有放回简单随机抽样两种方法。

不放回简单随机抽样比放回抽样有更低的抽样误差。

(2) 优缺点:

优点: 简单随机抽样是最基本的随机抽样方法, 操作简单, 且每个单位的入样概率相同, 因而样本估计量形式也比较简单。

缺点: 没有利用抽样框中更多的辅助信息, 所以用样本统计量估计总体参数的效率受到影响; 同时在简单随机抽样条件下, 样本的分布可能十分分散, 增加了调查过程中的费用和时间。

(3) 适用条件:

- ①抽样框中没有更多可以利用的辅助信息;
- ②调查对象分布的范围不广阔;
- ③个体之间的差异不是很大。

## 7、分层抽样 (高频考点、特别关注)

(1) 概念:

分层抽样是指先按照某种规则把总体分为不同的层, 然后在不同的层内独立、随机地抽取样本, 这样所得到的样本称为分层样本。如果每层中的抽样都是简单随机抽样, 则称为分层随机抽样。

(2) 特点:

- ①分层抽样不仅可以估计总体参数, 同时也可以估计各层的参数。
- ②便于抽样工作的组织。
- ③每层都要抽取一定的样本单位, 这样样本在总体中分布比较均匀, 可以降低抽样误差。

分层抽样中, 样本量在各层中分配的方法可以归为两类: 等比例分配和不等比例分配。

(3) 应用条件

抽样框中有足够的辅助信息, 能够将总体单位按某种标准划分到各层之中, 实现在同一层内, 各单位之间的差异尽可能地小, 不同层之间各单位的差异尽可能地大。

## 8、系统抽样 (高频考点、特别关注)

(1) 概念

系统抽样是指先将总体中的所有单位按一定顺序排列, 在规定范围内随机抽取一个初始单元, 然后按事先规定的规则抽取其他样本单元。最简单的系统抽样是等距抽样。

(2) 优缺点优点:

- ①操作简便, 只需要随机确定起始单位, 整个样本就自然确定了。
- ②对抽样框的要求也比较简单, 只要求总体单位按一定顺序排列, 而不一定是一份具体的名录清单。

缺点: 方差估计比较复杂, 给计算抽样误差带来一定困难。

(3) 系统抽样的估计效果与总体单位排列顺序有关。如果排列顺序与调查内容没有联系，称为按无关标识排列，这时系统抽样估计与简单随机抽样估计效率相仿。如果排列顺序与调查内容有关，称为按有关标识排列。

按有关标识排列的系统抽样精度一般比简单随机抽样的精度高。

## 9、整群抽样（高频考点、特别关注）

### (1) 概念

整群抽样是将总体中所有的基本单位按照一定规则划分为互不重叠的群，抽样时直接抽取群，对抽中的群调查其全部的基本单位，对没有抽中的群则不进行调查。

### (2) 优缺点：

①实施调查方便，可以节省费用和时间

②抽样框编制得以简化

缺点：由于抽取的样本单位比较集中，群内各单位之间存在相似性，差异比较小，而群与群之间的差别往往比较大，使得整群抽样的抽样误差比较大。

### (3) 适用

整群抽样特别适合于对某些特殊群结构进行调查。

## 10、多阶段抽样

### (1) 概念

在大规模抽样调查中，一次抽取到最终样本单位是很难实现的，往往需要经过二个或二个以上阶段才能抽到终样本单位，这就是多阶段抽样方法。

### (2) 在大范围抽样调查中采用多阶段抽样的原因

①在大范围抽样调查中，往往没有包括所有总体单位的抽样框，或者编制这样的抽样框十分困难。多阶段抽样是分阶段进行的，抽样框也可以分级进行准备。

②多阶段抽样是在中选单位中的再抽选，这样就使样本的分布相对集中，从而可以节省调查中的人力和财力。

### (3) 缺点

多阶段的抽样设计比较复杂，不仅涉及如何划分阶段，还包括在每个阶段上应当抽取多大样本量，以及每个阶段的抽样方法。此外，多阶段的抽样误差计算也比较复杂。

## 11、估计量的性质（高频考点）

### (1) 估计量的无偏性

对于不放回简单随机抽样，所有可能的样本均值取值的平均值总是等于总体均值，这就是样本均值估计量的无偏性。

### (2) 估计量的有效性

### (3) 估计量的一致性

随着样本量的增大，估计量的值如果稳定于总体参数的真值，这个估计量就有一致性。

## 12、抽样误差的估计（高频考点、特别关注）

①抽样误差与总体分布有关，总体单位值之间差异越大，则总体方差越大，抽样误差越大。

②抽样误差与样本量 $n$ 有关，在其他条件相同情况下，样本量越大，抽样误差就越小。

③抽样误差与抽样方式和估计量的选择也有关系。

利用有效辅助信息的估计量也可以有效地减少抽样误差。

13、样本量的影响因素

- (1) 调查的精度，同向。调查的精度是指用样本数据对总体进行估计时可以接受的误差水平。
- (2) 总体的离散程度，同向。
- (3) 总体的规模，对于小规模总体，总体规模越大，样本也要随之增大。
- (4) 无回答情况，若无回答率较高，样本要大一些。
- (5) 经费的制约，样本量是经费和精度的折中。
- (6) 调查的限定时间及实施调查的人力资源。

第二十六章回归分析

1、相关分析与回归分析的关系（特别关注）

|              |                                                 |                                                                                          |
|--------------|-------------------------------------------------|------------------------------------------------------------------------------------------|
| 定义           | 所谓回归分析就是根据相关关系的具体形态，选择一个合适的数学模型，来近似地表达变量间的依赖关系。 |                                                                                          |
| 相关分析与回归分析的关系 | 联系                                              | 它们不仅具有共同的研究对象，而且在具体应用时，常常必须互相补充。相关分析需要依靠回归分析来表明现象数量相关的具体形式，而回归分析则需要依靠相关分析来表明现象数量变化的相关程度。 |

|                                                                               |    |                            |                                                                                             |
|-------------------------------------------------------------------------------|----|----------------------------|---------------------------------------------------------------------------------------------|
| 相关分析<br>与回归分<br>析<br>的关系                                                      | 区别 | 相关分析和回归分析在研究目的和研究方法上有明显区别： |                                                                                             |
|                                                                               |    | 相关<br>分析                   | 研究变量之间相关的方向和相关程度。                                                                           |
|                                                                               |    | 回归分<br>析                   | 研究变量之间相互关系的具体形式，对具有相关关系的变量之间的数量联系进行测定，确定一个相关的数学方程式，根据这个数学方程式可以从已知量来推测未知量，从而为估算和预测提供了一个重要方法。 |
| 进行回归分析时，首先确定因变量和自变量；<br>被预测或被解释的变量称为因变量，一般用Y表示；<br>用来预测或解释因变量的变量称为自变量，一般用X表示。 |    |                            |                                                                                             |

2、一元线性回归模型（高频考点、特别关注）

一元线性回归是描述两个变量之间相关关系的最简单的回归模型。

回归模型可以用描述因变量Y如何依赖自变量X和误差项  $\epsilon$  的方程来表示。只涉及一个自变量的一元线性回归模型可以表示为：

$$Y = \beta_0 + \beta_1 X + \epsilon$$

式中  $\beta_0$  和  $\beta_1$  为模型的参数

y是x的线性函数( $\beta_0 + \beta_1 X$ )加上误差项  $\epsilon$ 。

$\beta_0 + \beta_1 X$ 反映了由于x的变化而引起的y的变化；

误差项  $\varepsilon$  是随机变量，反映了除  $x$  和  $y$  之间的线性关系之外的随机因素对  $y$  的影响，是不能由  $x$  和  $y$  之间的线性关系所解释的  $Y$  的变异性。

### 3、回归模型的拟合效果分析

（1）一般情况下，使用估计的回归方程之前，需要对模型进行检验：

- ①结合经济理论和经验分析回归系数的经济含义是否合理；
- ②分析估计的模型对数据的拟合效果如何；
- ③对模型进行假设检验。

（2）决定系数——一元线性回归模型拟合效果的一种测度方法（高频考点、特别关注）

决定系数，也称为  $R^2$

决定系数的取值在 0 到 1 之间，大体上说明了回归模型所能解释的因变量占因变量总变化的比例。

- ①决定系数  $R^2$  越高，越接近于 1，模型的拟合效果就越好，即模型解释因变量的能力越强。否则，越接近于 0，回归直线拟合效果越差。
- ②决定系数为 1，所有观测点都落在回归直线上，说明回归直线可以解释因变量的所有变化。
- ③决定系数为 0，说明回归直线无法解释因变量的变化，因变量的变化与自变量无关。

## 第二十七章时间序列分析

### 1、时间序列含义和构成要素：

**时间序列也称动态数列**，是将某一统计指标在各个不同时间上的数值按时间先后顺序编制形成的序列。时间序列的构成要素：

- （1）被研究现象**所属时间**；
- （2）反映该现象一定时间条件下数量特征的**指标值**。

### 2、时间序列的分类（高频考点）：

时间序列按照其构成要素中统计指标值的表现形式，分为：

- （1）绝对数时间序列：统计指标值是绝对数。根据指标值的时间特点又分为：

**时期序列**：每一指标值反映现象在一定时期内发展的结果。即过程总量。如某年的国内生产总值是一年中各月数加在一起的结果。

**时点序列**：每一指标值反映现象在一定时点上的瞬间水平。如全国人口数。

- （2）相对数时间序列：统计指标值是相对数。如男、女人口的比重。
- （3）平均数时间序列：统计指标值是平均数，如人均国内生产总值。

### 3、时间序列的水平分析（高频考点、特别关注）

时间序列的水平分析包括：发展水平、平均发展水平、增长量和平均增长量。

- （1）**发展水平**，是时间序列中对应于具体时间的指标数值。
- （2）**平均发展水平**

也称序时平均数或动态平均数，是对时间序列中各时期发展水平计算的平均数，它可以概括性描述现象在一段时期内所达到的一般水平。

#### ①绝对数时间序列序时平均数的计算

| 序列   |                   | 具体类别     | 平均数的计算                                                                                                                                     |
|------|-------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------|
| 时期序列 |                   |          |                                                                                                                                            |
| 时点序列 | 连续时点<br>(以天为时间单位) | 逐日登记逐日排列 | 简单算术平均数<br>$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{\sum_{i=1}^n y_i}{n}$                                                      |
|      |                   | 指标值变动才登记 | 加权算术平均数<br>$\bar{y} = \frac{y_1 f_1 + y_2 f_2 + \dots + y_n f_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^n y_i f_i}{\sum_{i=1}^n f_i}$ |

| 序列   |      | 具体类别    | 平均数的计算                                                                                                                                                           |
|------|------|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 时点序列 | 间断时点 | 间隔时间相等  | 两次平均：均为简单算术平均<br>$\bar{y} = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \dots + \frac{y_{n-1} + y_n}{2}}{n - 1}$                                           |
|      |      | 间隔时间不相等 | 两次平均：第一次简单算术平均；第二次加权算术平均<br>$\bar{y} = \frac{\frac{y_1 + y_2}{2} f_1 + \frac{y_2 + y_3}{2} f_2 + \dots + \frac{y_{n-1} + y_n}{2} f_{n-1}}{\sum_{i=1}^{n-1} f_i}$ |

【例-单选题】某企业职工人数资料（单位：人）如下：

| 时间   | 3月31日 | 4月30日 | 5月31日 | 6月30日 |
|------|-------|-------|-------|-------|
| 职工人数 | 2000  | 2500  | 2300  | 2200  |

该企业3~6月份平均职工人数为( )。

- A. 2350人
- B. 2200人
- C. 2250人
- D. 2300人

答案：D

解析：属于间断时点指标，每次登记的间隔期是1个月，所以：

$$\bar{y} = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \dots + \frac{y_{n-1} + y_n}{2}}{n - 1}$$

$$\bar{y} = \frac{\frac{2000 + 2500}{2} + \frac{2500 + 2300}{2} + \frac{2300 + 2200}{2}}{3} = 2300$$

## ②相对数或平均数时间序列序时平均数的计算

相对数或平均数时间序列是派生数列，相对数或平均数通常是由两个绝对数对比形成的。

计算思路：分别求出分子指标和分母指标时间序列的序时

平均数，然后再进行对比，用公式表示如下：

$$\bar{y} = \frac{\bar{a}}{\bar{b}}$$

## （3）增长量：（高频考点）

报告期发展水平与基期发展水平之差，反映报告期比基期增加（减少）的绝对数量。

用公式表示为：增长量 = 报告期水平 - 基期水平

①逐期增长量：报告期水平与上一期水平之差。

②累计增长量：报告期水平与某一固定时期（通常是时间序列最初水平）水平之差。

【注】同一时间序列中，累计增长量等于相应时期逐期增长量之和。

## （4）平均增长量（高频考点）

平均增长量是时间序列中逐期增长量的序时平均数，它表明现象在一定时段内平均每期增加（减少）的数量。其计算公式为：

$$\text{平均增长量} = \frac{\text{逐期增长量的合计}}{\text{逐期增长量的个数}} = \frac{\text{累计增长量}}{\text{时间序列项数} - 1}$$

## 4、时间序列的速度分析（高频考点、特别关注）

时间序列的速度分析包括：发展速度、增长速度、平均发展速度和平均增长速度。

## 5、发展速度（高频考点、特别关注）

是以相对数形式表示的两个不同时期发展水平的比值，表明报告期水平已发展到基期水平的几分之几或若干倍。

$$\text{发展速度} = \frac{\text{报告期水平}}{\text{基期水平}}$$

由于基期选择的不同，发展速度有定基与环比之分。

(1) 定基发展速度是报告期水平与某一固定时期水平（通常是最初水平）的比值，

用  $a_i$  表示，

$$\text{定基发展速度 } a_i = \frac{\text{报告期水平 } y_i}{\text{最初水平 } y_0}$$

(2) 环比发展速度是报告期水平与其前一期水平的比值，

用  $b_i$  表示，

$$\text{环比发展速度 } b_i = \frac{\text{报告期水平 } y_i}{\text{报告期前一期水平 } y_{i-1}}$$

(3) 定基发展速度与环比发展速度之间的关系

① 定基发展速度等于相应时期内各环比发展速度的连乘积——“定基积”

$$\begin{aligned} \text{推导：定基发展速度 } \frac{y_n}{y_0} &= \frac{y_1}{y_0} \times \frac{y_2}{y_1} \times \frac{y_3}{y_2} \times \dots \times \frac{y_n}{y_{n-1}} \\ &= \text{各环比发展速度的连乘积} \end{aligned}$$

② 两个相邻时期定基发展速度的比率等于相应时期的环比发展速度——“环比比”

$$\begin{aligned} \text{推导：相邻时期定基发展速度的比率 } & \frac{\frac{y_n}{y_0}}{\frac{y_{n-1}}{y_0}} \\ &= \frac{y_n}{y_{n-1}} = \text{相应时期的环比发展速度} \end{aligned}$$

6、增长速度（高频考点）

增长速度是报告期增长量与基期水平的比值，表明报告期比基期增长了百分之几或若干倍。

$$\text{增长速度} = \frac{\text{报告期增长量}}{\text{基期水平}} = \frac{\text{报告期水平} - \text{基期水平}}{\text{基期水平}} = \text{发展速度} - 1$$

(1) 定基增长速度：增长量是累计增长量

$$\text{定基增长速度} = \frac{\text{累计增长量}}{\text{基期水平}} = \frac{\text{报告期水平} - \text{基期水平}}{\text{基期水平}} = \text{定基发展速度} - 1$$

(2) 环比增长速度：增长量是逐期增长量

定基增长速度与环比增长速度之间的推算，必须通过定基发展速度与环比发展速度才能进行。

$$\text{环比增长速度} = \frac{\text{逐期增长量}}{\text{上期水平}} = \frac{\text{报告期水平} - \text{上期水平}}{\text{上期水平}} = \text{环比发展速度} - 1$$

## 7、平均发展速度与平均增长速度（高频考点、特别关注）

(1) 平均发展速度：反映现象在一定时期内逐期发展变化的一般程度。

平均发展速度是一定时期内各期环比发展速度的序时平均数。

目前计算平均发展速度通常采用几何平均法。

几何平均法也称水平法，其计算原理是：

一定时期内现象发展的总速度等于各期环比发展速度的连乘积，根据平均数的计算原理，就应当按连乘法，即几何平均数公式计算指标值的平均数。

$n$ 表示环比发展速度的时期数，也就是时间序列项数减1。

$$\text{平均发展速度} \quad \bar{b} = \sqrt[n]{\prod_{i=1}^n b_i} = \sqrt[n]{\text{环比发展速度连乘积}}$$

$$\text{或者：} \quad \bar{b} = \sqrt[n]{\frac{y_n}{y_0}} = \sqrt[n]{\text{定基发展速度}}$$

(2) 平均增长速度：反映现象在一定时期内逐期增长（降低）变化的一般程度。

(3) 平均发展速度与平均增长速度的关系：

平均增长速度 = 平均发展速度 - 1

## 8、速度的分析与应用（高频考点、特别关注）

在应用速度分析实际问题时，须防止误用乃至滥用的现象，应注意：

(1) 当时间序列中的指标值出现0或负数时，不宜计算速度。

(2) 速度指标的数值与基数的大小密切相关。

在环比增长速度时间序列中，各期的基数不同，因此，运用这一指标反映现象增长的快慢时，往往要结合水平指标的分析才能得出正确结论。

## 9、平滑预测法

平滑法的目的就是“消除”时间序列的不规则成分所引起的随机波动，所以被称为平滑法。平滑法适合于平稳时间序列的预测，即没有明显的趋势、循环和季节波动的时间序列。

### （1）移动平均法

移动平均法使用时间数列中最近k期数据值的平均数作为下一期的预测值，

其中，就是对时间序列的 $Y_t$ 预测结果；k为移动间隔（ $1 < k < t$ ）；

对于t+1期的简单移动平均预测值为：

$$F_{t+1} = \bar{Y}_t = \frac{Y_{t-k+1} + Y_{t-k+2} + \cdots + Y_{t-1} + Y_t}{k}$$

### （2）指数平滑法

指数平滑法是利用过去时间序列值的加权平均数作为预测值，即使得第t+1期的预测值等于第t期的实际观察值与第t期预测值的加权平均值。其基本计算公式为：

$$F_{t+1} = \alpha Y_t + (1 - \alpha) F_t$$

其中， $F_{t+1}$ 和 $F_t$ 分别为第t+1期和第t期的指数平滑预测值； $Y_t$ 为第t期的实际观察值； $\alpha$ 为平滑系数（即权重），取值范围为 $0 < \alpha < 1$ 。