

高一年级数学第 52 课时随机抽样（2）学习指南

一、学习目标

- （1）了解分层随机抽样的特点、适用范围、步骤和必要性；
- （2）掌握分层随机抽样中各层样本量比例分配的方法以及估计总体均值的方法；
- （3）通过实际情境对比，会选择恰当的抽样方法解决问题；
- （4）了解获取数据的基本途径.

二、学法指导

上节课我们学习了简单随机抽样，抽样方式简单、直观，是随机抽样理论中最基本的组织形式，它适用于小样本抽样，对于较大样本抽样，抽签法和随机数表法使用起来都比较麻烦，当总体分布不均匀时，简单随机抽样抽取的样本的代表性一般都不太好，总的来说简单随机抽样适用于总体量不太庞大，以及总体分布比较均匀，对于分布不均匀甚至数据差异明显的数据的较大样本该怎么进行抽样？

任务一：提出分层随机抽样的概念，归纳特点和步骤

问题 1 什么叫简单随机抽样？

一般地，设一个总体含有 N 个个体，从中逐个不放回地抽取 n ($1 \leq n \leq N$) 个个体作为样本,每次抽取时总体内的未进入样本的各个个体被抽到的概率都相等，就把这种抽样方法叫做简单随机抽样。

引例：小明想考察一下简单随机抽样的估计效果，他从树人中学医务室得到了高一年级学生 712 人身高的所有数据，其中男生 326 人，女生 386 人，计算出整个年级学生的平均身高为 165.0cm. 小明用简单随机抽样的方法，从这些数据中抽取了样本容量为 50 的样本 10 个，分别计算出平均值，如表 1 所示，

表 1

抽样序号	1	2	3	4	5	6	7	8	9	10
样本平均数	165.2	162.8	164.4	164.4	165.6	164.8	165.3	164.7	165.7	165.0

图像表示如图 1，其中图 1 中红色的线表示树人中学高一年级全体学生身高的平均值.

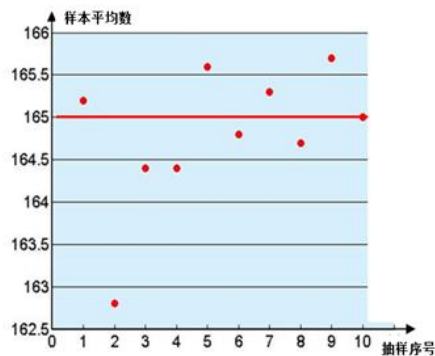


图 1

问题 2: 所抽 2 号样本平均值仅为 162.8，这样的样本是怎么产生的？

抽取的样本女生较多或者全部是女生，男生较少或者没有，出现"极端样本".

问题 3: 什么因素对身高产生了重大影响？为什么？

性别因素，男女生的身高差异较大.

问题 4: 在简单随机抽样中用样本均值估计总体均值,总体中的个体差异大小对估计效果会有什么影响？

个体差异越大，估计效果可能会越差，个体差异越小，估计效果可能会越好.

问题 5: 如何改进简单随机抽样，减少身高个体差异对估计效果的影响，使得抽取的样本更具代表性？

将高一学生分成男生和女生两个身高有明显差异的群体,对两个群里分别进行简单随机抽样再汇总成一个样本.

问题 6: 对男生和女生分别进行简单随机抽样，样本量应该如何分配？

$$\text{男生样本量} = \frac{\text{男生人数}}{\text{全体学生数}} \times \text{总样本量}$$

$$\text{女生样本量} = \frac{\text{女生人数}}{\text{全体学生数}} \times \text{总样本量}$$

注: 保证每个学生都有相同的可能性被抽到.

问题 7: 若抽取样本量为 50 的样本，男女生应该分别抽取多少人？

$$n_{\text{男}} = \frac{326}{712} \times 50 \approx 22.9$$

$$n_{\text{女}} = \frac{386}{712} \times 50 \approx 27.1$$

所以男女生应该分别抽取 23 人和 27 人

问题 8: 我们把上述抽样方式称为分层随机抽样，试归纳分层随机抽样的概念？

分层随机抽样: 按一个或多个变量把总体划分成若干个子总体，每个个体属于且仅属于一个子总体，在每个子总体中独立地进行简单随机抽样，再把所有子总体中抽取的样本合在一起作为总样本，这样的抽样方法称为**分层随机抽样**，每一个子总体称为**层**. 在分层随机抽样中，如果每层样本量都与层的大小成比例，那么称这种样本量的分配方式为**比例分配**.

特征: (1) 层与层不相交，分层原则是层内个体差异要小，层间个体差异要大.

(2) 按比例分配的分层随机抽样中，每个个体入样的概率相等.

(3) 在每层抽样时，可采用简单随机抽样的方法进行抽样.

适用范围: 由差异明显的几部分数据构成的较大样本

分层随机抽样步骤

- (1) 分层：将总体按某种特征或者多种特征分成若干不相交层；
- (2) 求比：计算各层的个体数与总体的个体数的比例；
- (3) 定数：按各层个体数与总体的个体数的比例确定各层应抽取的样本量；
- (4) 抽样：对每一层进行抽样（各层采用简单随机抽样），综合每层抽样，组成样本。

任务二：按比例分配分层随机抽样总样本均值估计

探究 1：引例中，用按比例分配分层随机抽样的方法抽取了一个容量为 50 的样本，观测数据（单位：cm）如下：

男生

173.0 174.0 166.0 172.0 170.0 165.0 165.0 168.0 164.0 173.0 172.0 173.0
175.0 168.0 170.0 172.0 176.0 175.0 168.0 173.0 167.0 170.0 175.0

女生

163.0 164.0 161.0 157.0 162.0 165.0 158.0 155.0 164.0 162.5 154.0 154.0
164.0 149.0 159.0 161.0 170.0 171.0 155.0 148.0 172.0 162.5 158.0 155.5
157.0 163.0 172.0

- (1) 估计树人中学高一年级男生、女生的平均身高。
- (2) 估计树人中学高一年级全体学生的平均身高。

解：(1) 经过计算，男生和女生身高的样本平均数分别为 170.6，160.6，所以估计高一年级男生的平均身高为 170.6 cm，高一年级女生的平均身高为 160.6 cm。

追问：这样估计为什么合理？

- (2) **方法一：**用抽取样本中的男女身高平均值和总体中男女的人数估计总体均值。

$$\bar{U} = \frac{170.6 \times 326 + 160.6 \times 386}{712} = 165.17 \approx 165.2$$

思考：能否用样本均值估计总体均值？为什么

方法二：用抽取样本中的男女身高平均值和样本中男女的人数（样本均值）估计总体均值。

$$\bar{w} = \frac{170.6 \times 23 + 160.6 \times 27}{50} = 165.2$$

$$\bar{w} = 170.6 \times \frac{23}{50} + 160.6 \times \frac{27}{50}, \quad \bar{U} = 170.6 \times \frac{326}{712} + 160.6 \times \frac{386}{712}$$

$$\text{因为是按比例抽样，有 } \frac{23}{50} \approx \frac{326}{712}, \frac{27}{50} \approx \frac{386}{712}$$

$$\text{所以 } \bar{w} \approx \bar{U}$$

小结：在按比例分配分层随机抽样中，可直接用样本平均数估计总体平均数。

思考：一般的，在按比例分配分层随机抽样中用样本均值估计总体均值，是否也可行？

在按比例分层随机抽样中，如果层数分为 2 层，第 1 层和第 2 层包含的个体数分别为 M 和 N ，抽取的样本量分别为 m 和 n 。用 $X_1, X_2, \dots, X_M$ 表示第 1 层各个体的变量值，用 $x_1, x_2, \dots, x_m$ 表示第 1 层样本的各个体的变量值，用 $Y_1, Y_2, \dots, Y_N$ 表示第 2 层各个体的变量值，用 $y_1, y_2, \dots, y_n$ 表示第 2 层样本的各个体的变量值。

第一层总体平均数和样本平均数为

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_M}{M} = \frac{1}{M} \sum_{i=1}^M X_i, \quad \bar{x} = \frac{x_1 + x_2 + \dots + x_m}{m} = \frac{1}{m} \sum_{i=1}^m x_i,$$

第二层总体平均数和样本平均数为

$$\bar{Y} = \frac{Y_1 + Y_2 + \dots + Y_N}{N} = \frac{1}{N} \sum_{i=1}^N Y_i, \quad \bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{1}{n} \sum_{i=1}^n y_i,$$

总体平均数和样本平均数

$$\bar{W} = \frac{\sum_{i=1}^M X_i + \sum_{i=1}^N Y_i}{M + N}, \quad \bar{w} = \frac{\sum_{i=1}^m x_i + \sum_{i=1}^n y_i}{m + n},$$

由于用第 1 层的样本平均数 $\bar{x}$ 可以估计第 1 层的总体平均数 $\bar{X}$ ，用第 2 层的样本平均数 $\bar{y}$ 可以估计第 2 层的总体平均数 $\bar{Y}$ ，因而我们可以用

$$\frac{M \times \bar{x} + N \times \bar{y}}{M + N} = \frac{M}{M + N} \bar{x} + \frac{N}{M + N} \bar{y}$$

估计总体平均数 $\bar{W}$ 。

在比例分配的分层随机抽样中有 $\frac{m}{M} = \frac{n}{N} = \frac{m+n}{M+N}$

可得 $\frac{M}{M+N} \bar{x} + \frac{N}{M+N} \bar{y} = \frac{m}{m+n} \bar{x} + \frac{n}{m+n} \bar{y} = \bar{w}$ 。

思考：以上结论是否能推广到三层、甚至更多层？

可以

探究 2 在引例中，采用按比例分配的分层随机抽样方法，从高一年级的学生中抽取了 10 个样本量为 50 的样本，计算出样本平均数如表 2，

表 2

	抽样序号									
	1	2	3	4	5	6	7	8	9	10
男生样本平均数	170.0	170.7	169.8	171.7	172.7	171.9	171.6	170.6	172.6	170.9
女生样本平均数	162.2	160.3	159.7	158.1	161.1	158.4	159.7	160.0	160.6	160.2
总样本平均数	165.8	165.1	164.3	164.3	166.4	164.6	165.2	164.9	166.1	165.1

对比引例中的简单随机抽样的结果，如图 2 所示（红色的线为整个年级学生的平均身高），你有什么发现？

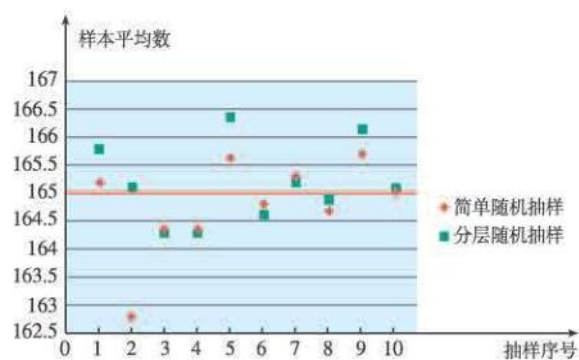


图 2

(1) 分层随机抽样的样本平均数围绕总体平均数波动，与简单随机抽样的结果比较，分层随机抽样并没有明显优于简单随机抽样。

(2) 相对而言，分层随机抽样的样本平均数波动幅度更均匀，简单随机抽样的样本平均数波动幅度更大，如第 2 个，即出现了比较“极端”的样本，而分层随机抽样几乎没有出现。

思考：为什么分层随机抽样没有明显好于简单随机抽样？

(1) 试验次数不够多，实验结果在随机性；

(2) 数据本身男生女生层内个体差异不是很小，分层不是非常合理。

说明：在个体差异较大的情形下，只要选取的分层变量合适，使得各层间差异明显、层内差异不大，分层随机抽样的效果一般都会好于简单随机抽样，也好于其他抽样方法。

任务四：分层随机抽样应用

例 1：下列问题中，采用怎样的抽样方法较为合理？

(1) 从 10 台电冰箱中抽取 3 台进行质量检查；

(2) 某学校有 160 名教职工，其中教师 120 名，行政人员 16 名，后勤人员 24 名，为了了解教职工对学校在校务公开方面的意见，拟抽取一个容量为 20 的样本。

解：(1) 简单随机抽样 (2) 分层随机抽样

思考：简单随机抽样和分层随机抽样有什么区别和联系？

类别	各自特点	相互联系	适用范围
简单随机抽样	从总体中逐个抽取	各层抽样采用简单随机抽样	总体中的个体数较少
分层随机抽样	将总体分成几层，分层进行抽取		由差异明显的几部分组成且样本量较大的总体

例 2：高二年级有男生 490 人，女生 510 人，张华按男生、女生进行分层，通过分层随机抽样的方法，得到男生、女生的平均身高分别为 170.2cm 和 160.8cm.

- (1) 若张华在各层中按比例分配样本，总样本容量为 100，那么在男生、女生中分别抽取多少人？请估计高二年级全体学生的平均身高.
- (2) 如果张华从男生、女生中抽取的样本量分别为 30 和 70，那么在这种情况下，如何估计高二年级学生的平均身高更合理？

解：(1) 设在男生、女生中分别抽取 m 名和 n 名，则

$$\begin{aligned} m &= 100 \times \frac{490}{1000} = 49 \\ n &= 100 \times \frac{510}{1000} = 51 \end{aligned}$$

由此可以估计高二年级的平均身高为

$$\frac{490}{1000} \times 170.2 + \frac{510}{1000} \times 160.8 \approx 165.4(cm)$$

或

$$\frac{49}{100} \times 170.2 + \frac{51}{100} \times 160.8 \approx 165.4(cm)$$

(2) 估计高二年级的平均身高为

$$\frac{490}{1000} \times 170.2 + \frac{510}{1000} \times 160.8 \approx 165.4(cm)$$

追问：为什么用样本平均值不能较好的估计总体均值？

抽取样本不具较好的代表性，不能较好的反映总体信息.

任务五：了解获取数据的基本途径

一、获取数据的基本途径

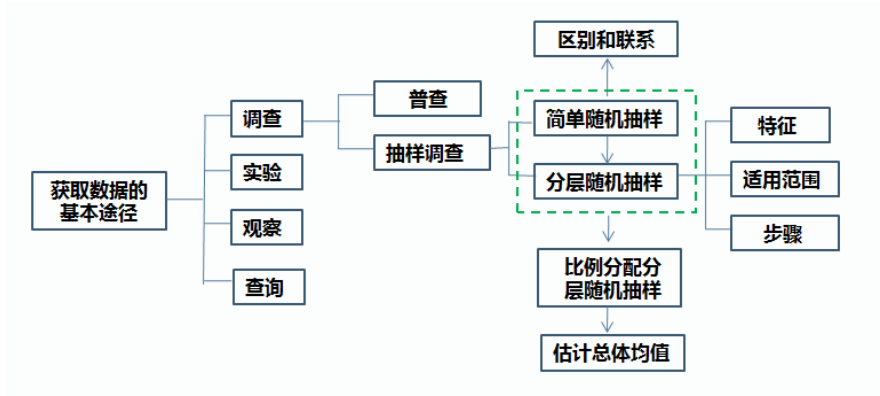
获取数据的基本途径	特征	举例
通过调查获取数据	对有限总体一般通过抽样调查或普查的方法获取的数据	人口总数、城乡就业状况、农村贫困人口脱贫状况、生态环境改善状况、青少年受教育状况、高中生近视的比例、产品合格率、高中生日平均上网时间等数据
通过试验获取数据	没有现存可以查询的数据	判断研制的新药是否有效、培育的小麦新品种是否具有更高的产量，特种钢、轮胎的配方和产品质量等数据
通过观察获取	主要是自然现象产生的数据	地震、降水、大气污染、宇宙射线等数据

数据		
通过查询获得数据	众多专家研究过收集和存储的数据	年度统计公报、经济普查公报、人口普查公报、农业普查公报、《中国统计摘要》、《中国统计年鉴》、国家统计局等能查询到的数据

二、部分数据获取查询网址

国家统计局: <http://www.stats.gov.cn/> 中经数据: <https://ceidata.cei.cn/>
 UCI 数据库: <https://archive.ics.uci.edu/ml/index.php> 百度学术 <http://xueshu.baidu.com/>
 谷粉学术 <https://wenku.baidu.com/> 中国知网: <https://www.cnki.net/>
 万方数据: <http://www.wanfangdata.com.cn/index.html> 《科学》杂志 <https://www.sciencemag.org/>
 《自然》杂志 <https://www.nature.com/> springerlink 数据库 <https://link.springer.com/>

回顾本节知识脉络



数据获取篇

1 国内统计数据

1.1 综合类统计数据

综合类统计年鉴 各级统计局每年 9 月份左右出版（现在有许多是电子版）

国民经济和社会发展统计公报 各级统计局每年 2 月份以后发布（网上发布）

1.2 行业类统计数据

行业统计年鉴 由政府主管部门组编 出版时间不定

行业统计公报 由政府主管部门 4 月份以后发布（网上发布）

行业年鉴 各行业协会组编、每年 8 月份以后出版（知网数据库有电子版）

2 国际组织发布的数据

2.1 综合性组织发布的数据

联合国、世界银行、世界贸易组织、国际货币基金组织等

2.2 行业性组织发布的数据

国际海事组织、国际港口协会、国际航运公会、世界航运协会、美国港口协会、欧洲港口协会等

3 研究类数据

3.1 科研机构发布的数据

美国西北大学运输研究中心、英国利兹大学交通研究所、集装箱化与多式联运研究所等

中国科学院、中国工程院、中国社会科学院、国务院发展研究中心、国家发改委综合运输研究所、上海市政府发展研究中心、上海市城市综合交通规划研究所

3.2 咨询机构发布的数据

综合类：兰德公司、麦肯锡、埃森哲、罗兰贝格、科尔尼、普华永道、毕马威等

交运类：德鲁里、克拉克森、劳氏海运信息服务公司、上海国际航运研究中心

4 各类公司发布的数据

4.1 工商业公司发布的年报

韩国三星重工、日本三菱重工、利勃海尔集团、斯堪尼亚集团、鹿特丹港口集团、新加坡 PSA 国际港务集团、马士基、地中海、达飞、东方海外等

中船重工、中船集团、振华重工、中国重汽、中集集团、上港集团、中远集运、中海集运等

4.2 证券公司发布的行业分析报告

中信证券、兴业证券、民生证券、华泰证券、海通证券等；

国泰君安、申银万国、中金公司、莫尼塔等；

慧博投研、渐飞数据伴侣等

5 特别说明

5.1 数据时间

一般国内年鉴里的数据为上一年的数据，国际年鉴里的数据为前 2 年的数据；也就是国内数据滞后期为 1 年，国际数据滞后期为 2 年。

5.2 获取渠道

图书馆纸质图书、图书馆数据库、百度文库、爱问共享、道客巴巴等 相关组织、公司网站