



全国统计教材编审委员会“十二五”规划教材

应用抽样技术： 学习指导与习题解答

新 编

杨贵军 孟杰 编著

 中国统计出版社
China Statistics Press



全国统计教材编审委员会“十二五”规划教材

应用抽样技术： 学习指导与习题解答

新 编

杨贵军 孟杰 编著

 中国统计出版社
China Statistics Press

图书在版编目(CIP)数据

应用抽样技术：学习指导与习题解答 / 杨贵军，孟杰编著.

—— 北京：中国统计出版社，2017.7

全国统计教材编审委员会“十二五”规划教材

ISBN 978—7—5037—8136—0

I. ①应… II. ①杨… ②孟… III. ①抽样调查统计
— 高等学校—教材 IV. ①C811

中国版本图书馆 CIP 数据核字(2017)第 090672 号

应用抽样技术:学习指导与习题解答

作 者/杨贵军 孟杰

责任编辑/姜 洋

封面设计/上智博文

出版发行/中国统计出版社

通信地址/北京市丰台区西三环南路甲 6 号 邮政编码/100073

电 话/邮购(010)63376909 书店(010)68783171

网 址/www.zgtjcbbs.com

印 刷/河北鑫兆源印刷有限责任公司

经 销/新华书店

开 本/710×1000mm 1/16

字 数/230 千字

印 张/13.5

版 别/2017 年 7 月第 1 版

版 次/2017 年 7 月第 1 次印刷

定 价/29.00 元

版权所有。未经许可，本书的任何部分不得以任何方式在世界任何地区
以任何文字翻印、仿制或转载。

中国统计版图书，如有印装错误，本社发行部负责调换。

国家统计局

全国统计教材编审委员会

顾 问 袁 卫 冯士雍 吴喜之 方积乾

王吉利 庞 皓 李子奈

主 任 徐一帆

副主任 严建辉 夏荣坡 邱 东 施建军

耿 直 徐勇勇

委 员(按姓氏笔划排序)

丁立宏	万崇华	马 骏	毛有丰	王兆军
王佐仁	王振龙	王惠文	丘京南	史代敏
龙 玲	刘建平	刘俊昌	向书坚	孙秋碧
朱 胜	朱仲义	许 鹏	余华银	张小斐
张仲梁	张忠占	李 康	李兴绪	李宝瑜
李金昌	李朝鲜	杨 虎	杨汭华	杨映霜
汪荣明	肖红叶	苏为华	陈 峰	陈相成
房祥忠	林金官	罗良清	郑 明	柯惠新
柳 青	胡太忠	贺 佳	赵彦云	赵耐青
凌 亢	唐年胜	徐天和	徐国祥	郭建华
崔恒建	傅德印	景学安	曾五一	程维虎
蒋 萍	潘 璠	颜 虹		

出版说明

全国统计教材编审委员会是国家统计局领导下的、全国统计教材建设工作的最高指导机构和咨询机构,自1988年成立以来,分别组织编写和出版了“七五”至“十一五”全国统计规划教材。

“十二五”时期,是我国全面实施素质教育,全面提高高等教育质量,深化教育体制改革,推动教育事业科学发展,提高教育现代化水平的时期。“十二五”伊始,统计学迎来了历史性的重大变革和飞跃。2011年2月,在国务院学位委员会第28次会议通过的新的《学位授予和人才培养学科目录(2011)》(以下简称“学科目录”)中,统计学从数学和经济学中独立出来,成为一级学科。这一变革和飞跃将对中国统计教育事业产生巨大而深远的影响,中国统计教育事业将在“十二五”时期发生积极变化。

正是在这一背景下,全国统计教材编审委员会制定了《“十二五”全国统计教材建设规划》(以下简称“规划”)。根据“学科目录”在统计学下设有数理统计学,社会经济统计学,生物卫生统计学,金融统计、风险管理与精算学,应用统计5个二级学科的构架,“规划”对“十二五”全国统计规划教材建设作了全面部署,具有以下特点:

第一,打破以往统计规划教材出版学科单一的格局。全面发展数理统计学,社会经济统计学,生物卫生统计学,金融统计、风险管理与精算学,应用统计5个二级学科规划教材的出版,使“十二五”全国统计规划教材涵盖5个二级学科,形成学科全面并平衡发展的出版局面。

第二,打破以往统计规划教材出版层次单一的格局。在编写出版好各学科本科生教材的基础上,对研究生教材出版进行深入研究,出版一批高水平高层次的研究生教材,为我国研究生教育、尤其是应用统计研究生教育提供教学服务。同时,积极重视统计专科教材出版,联合各专科院校,组织编写和出版适应统计专科教学和学习的优秀教材。

第三,打破以往统计规划教材出版品种单一的格局。鼓励内容创

新,联系统计实践,具有教学内容和教学方法特色的、各高校自编的相同内容选题的精品教材出版,促进统计教学向创新性、创造性和多样性发展。

第四,重视非统计专业的统计教材出版。探讨对非统计专业学生的统计教学问题,为非统计专业学生组织编写和出版概念准确、叙述简练、深入浅出、表达方式活泼、练习题贴近社会生活的统计教材,使统计思想和统计理念深入非统计专业学生,以达到统计教学的最大效果。

第五,重视配合教师教学使用的电子课件和辅助学生学习使用的电子产品的配套出版,促进高校统计教学电子化建设,以期最后能形成系统,提高统计教育现代化水平。

第六,重视对已经出版的统计规划教材的培育和提高,本着去粗存精、去旧加新、与时俱进的原则,继续优化已经出版的统计教材的内容和写作,强化配套课件和习题解答,使它们成为精品,最后锤炼成为经典。

“十二五”期间,编审委员会将本着“重质量,求创新,出精品,育经典”的宗旨,组织我国统计教育界专家学者,编写和编辑出版好本轮教材。本轮教材出版后,将能够形成学科齐全、层次分明、品种多样、配套系统的高质量立体式结构,使我国统计规划教材建设再上新台阶,这将对推动我国统计教育和统计教材改革,推动我国统计教育事业科学发展,提高我国统计教育现代化水平产生积极意义。

让教师的教学和学生的学习事半功倍,并使学生在毕业之后能够学以致用用的统计教材,是本轮教材的追求。编审委员会将努力使本轮教材好教、好学、好用,尽力使它们在内容上和形式上都向国外先进统计教材看齐。限于水平和经验,在教材的编写和编辑出版过程中仍会有不足,恳请广大师生和社会读者提出批评和建议,我们将虚心接受,并诚挚感谢!

国家统计局
全国统计教材编审委员会
2012年7月

序 言

随着大数据应用的不断扩张,抽样技术受到越来越广泛的关注。统计学者并不认为抽样技术将会被淘汰,反而觉得其更加重要。大数据为抽样技术提供了更多应用领域。另一方面,抽样技术也需要不断创新,以便更好地适应大数据时代发展。抽样技术作为一门本科的统计学课程,主要内容还应该集中于几种常用的抽样方法,让读者能更好地理解抽样技术的思想。同时,掌握这些常用的抽样方法也是进一步学习前沿抽样方法的基础。本书集中在这几种常用的抽样方法,与全国统计教材编审委员会“十二五”规划教材《应用抽样技术》教材配套。

《应用抽样技术:学习指导与习题解答》按教材的章节配置,共分为七个章节,主要包括抽样技术的基本概念、简单随机抽样、分层随机抽样、比率估计和回归估计、等距抽样、单级整群抽样、两级抽样。每个章节的结构都按照知识要点、习题解答和习题解答的 R 语言代码等内容进行编排。这样可以与教材进行合理搭配,便于读者更好地把握每章内容的教学重点和解析每章的习题,有效地学习和理解抽样技术的基础知识。

在书稿完成过程中,兼顾以下的两方面。

(1)强调了习题解答的严谨性。尽管抽样技术应用需要有一定的调查经验,不能把书本知识直接应用于调查实践,但是本书还是希望读者能够较严谨地回答每道习题,促使读者能够从理论上理解抽样方法的特点和应用条件,将对抽样技术的正确理解作为抽样技术应用的基础,更好地把握抽样技术应用中的方法运用合理性和结论可靠性。

(2)注重习题解答的 R 语言实现。在抽样调查实践中,出现了调查数据分析的很多手工计算技术,但在习题解答中采用较少。随着计算机普及,读者也希望其能够代替繁琐的计算,而将重点集中于学习抽样方法的性质。为了满足读者这方面需求,本书提供了习题解答的 R 语言代码,便于模仿,便于理解。

本书是由杨贵军和孟杰共同编著完成。在书稿完成过程中,多个年级的在读硕士研究生和在读博士研究生参与了书稿的整理、修订与校对。作者在此感谢天津财经大学中国经济统计研究中心和统计学系的支持。感谢中国统计出版社的大力支持和协助。感谢国家自然科学基金(11471239)和国家社会科学基金(17CTJ002)资助。本书的部分习题借鉴了相关参考书中的例题或习题,在此特向有关作者表示感谢。在书稿完成过程中,尽管我们付出了最大的努力,但书中仍会有一些遗憾。对于书中的不足,恳请各位专家和读者提出宝贵意见。

作 者

2017 年 3 月

目 录

第 1 章 绪论	(1)
1.1 基本概念和定义	(1)
1.2 常用的总体标志和样本统计量	(3)
1.3 习题解答	(4)
第 2 章 简单随机抽样	(13)
2.1 简单随机抽样的基本概念和定义	(13)
2.2 简单随机抽样的估计量	(14)
2.3 简单随机抽样情形下的样本量确定	(19)
2.4 习题解答	(21)
第 3 章 分层随机抽样	(44)
3.1 分层随机抽样的基本概念和定义	(44)
3.2 分层随机抽样的估计量	(45)
3.3 分层随机抽样情形下的样本量确定	(48)
3.4 习题解答	(51)
第 4 章 比率估计量和回归估计量	(89)
4.1 比率估计量和回归估计量的基本概念和定义	(89)
4.2 比率估计量和回归估计量及其性质	(90)
4.3 习题解答	(100)

第 5 章 等距抽样	(140)
5.1 等距抽样的基本概念和定义	(140)
5.2 等距抽样的估计量及其性质	(141)
5.3 习题解答	(143)
第 6 章 单级整群抽样	(150)
6.1 单级整群抽样的基本概念和定义	(150)
6.2 单级整群抽样的估计量	(152)
6.3 单级整群抽样的样本量确定	(160)
6.4 习题解答	(161)
第 7 章 两级抽样	(182)
7.1 两级抽样的基本概念和定义	(182)
7.2 两级抽样的估计量	(183)
7.3 习题解答	(190)

第1章 绪论

1.1 基本概念和定义

1. 调查:常常用于描述通过观察、访谈和问卷等方法收集特定目标的数据或信息,并进行汇总和分析的过程。

2. 全面调查:全面调查是对被调查对象总体中的全部单元进行调查。普查是全面调查的一种形式。

3. 非全面调查:对总体中少数而不是全部的单元进行调查。非全面调查包括两种抽样方式,概率抽样和非概率抽样。

4. 概率抽样:事前确定每个单元被抽中的概率,再按照给定的概率进行抽样。当把概率抽样方法应用于同一具体总体时,能够被抽取的所有可能的样本都事前能够确定;并且每个样本都有已知的并且大于零的被抽中概率;这个概率等于用概率抽样方法随机抽取到该样本的概率。

5. 非概率抽样:不满足概率抽样的要求,常常是有意抽取具有某种特征单元的抽样方法。主要的非概率抽样有目的抽样、方便选样、定额抽样。

6. 目的抽样:也称为判断抽样,是调查组织者根据对调查目标的了解,挑选标志值较“适中”的单元,对总体进行估计。

7. 方便选样:采用某种方便方式从调查总体中选择单元。

8. 定额抽样:把单元按照类型分成若干组,从每组中用方便选样方式抽取一定数量的单元。

9. 描述性调查和分析性调查:描述性调查的目的是为了获取调查总体的某些信息;分析性调查是为了对总体内不同类的个体进行比较,以便发现它们之间的差异,分析或检验产生这些差异的原因。描述性调查和分析性调查之间没有明显区分,很多抽样调查都能达到这两个目的。

10. 抽样误差和非抽样误差:抽样误差是由于利用部分单元的数据推断总体数量特征所引起的误差。抽样误差是随机的,不是错误。除了抽样误差之外,抽样调查的其他因素造成的误差属于非抽样误差。理论上,非抽样误差是可以避免的。

非抽样误差分为三类,与抽样框有关的误差、调查数据无回答的误差、测量和记录观测数据的计量误差。与抽样框有关的误差主要是指抽样框所描述的总体与所研究的总体不一致。调查数据无回答的误差是指单元数据缺失所造成的误差。测量和记录观测数据的计量误差是指总体单位计量不准确所造成的误差。

11. 总体:是由某一统计调查任务所规定的,在时间、空间及若干其他标志上具有共同性质的客观存在的全部(有限个)单元(个体)组成的整体,即总体是单元的集合。总体是随机抽取的全部可能结果,即总体中每个单元的标志值,总体是标志值的集合。

12. 基本单元和群单元:基本单元也称总体基本单元、个体,是与总体相对应的概念。基本单元是总体中承担调查标志的个体。由基本单元组成的集合叫群单元。

13. 抽样单元:是实施抽样的单元。抽样单元可能是基本单元,也可能是群单元。

14. 个体抽样和整群抽样:在抽样实践中,以基本单元为抽样单元的抽样称为个体抽样;以群单元为抽样单元的抽样则称为整群抽样。

15. 抽样框:是全部抽样单元的完整列表,包括编号和对应的抽样单元名称。在抽样框中,编号与抽样单元一一对应。同一编号在抽样框中只能出现一次,不能出现多次,即同一编号不能对应两个及两个以上不同的抽样单元。

16. 目标总体和被抽样总体:目标总体是作为调查目标的总体。被抽样总体,也称作业总体,是抽样框所描述的总体。

17. 有限总体:是包含有限数量的基本单元。抽样调查的总体是有限总体,包含有限数量的基本单元。总体是由确定的统计调查任务规定的,在时间、空间及若干其他标志上具有共同性质的全部客观存在的有限数量的基本单元的全体。

18. 有限总体概率抽样:从有限总体中抽取单元,总体的每个单元都可能被抽取,并且每个单元都有确定的,大于零的被抽取概率。

19. 样本:从总体中用某种抽样方法抽取的 n 个抽样单元。 n 称为样本量,也称为样本容量、样本单元数。样本的抽样单元也叫样本单元。

20. 样本统计量:简称为统计量,是样本的函数,但不包含未知参数。不同的样本函数为不同的样本统计量。

21. 估计:对随机抽取的样本单元进行现场调查,获得样本观察值,按照估计量的计算公式,计算估计量的观察值,并作为总体标志的估计值,这个过程称为估计。估计总体标志的统计量为估计量。估计结果的常用评价标准有无偏性、有效性和最小均方误差。

22. 偏差、有偏估计量和无偏估计量:假设估计量的数学期望存在,偏差为估计量的期望与总体标志真值的差值。总体标志真值用 θ 表示,估计量用 $\hat{\theta}$ 表示。估计量的偏差为 $B(\hat{\theta}) = |E(\hat{\theta}) - \theta|$ 。如果 $E(\hat{\theta}) \neq \theta$,即估计量 $\hat{\theta}$ 的期望不等于总体标志真值 θ ,则 $B(\hat{\theta}) \neq 0$, $\hat{\theta}$ 称为总体标志 θ 的有偏估计量。如果 $E(\hat{\theta}) = \theta$,即估计量 $\hat{\theta}$ 的期望等于总体标志 θ 真值,则 $B(\hat{\theta}) = 0$, $\hat{\theta}$ 称为总体标志 θ 的无偏估计量。

23. 均方误差:估计量 $\hat{\theta}$ 的均方误差为 $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$ 。均方误差的平方根 $\sqrt{MSE(\hat{\theta})}$ 称为标准误。均方误差包括两部分,估计量 $\hat{\theta}$ 的方差 $E[\hat{\theta} - E(\hat{\theta})]^2$;估计量 $\hat{\theta}$ 的偏差平方 $[E(\hat{\theta}) - \theta]^2$ 。估计量的均方误差是估计量的方差和偏差的综合评价。

24. 点估计:就是构造一个适当的统计量 $\hat{\theta}(y_1, \dots, y_n)$,用统计量的观察值估计总体标志 θ 。这里称 $\hat{\theta}(y_1, \dots, y_n)$ 为 θ 的估计量,称 $\hat{\theta}(y_1, \dots, y_n)$ 的观察值为总体参数 θ 的估计值。在不致混淆的情况下,统称估计量 $\hat{\theta}(y_1, \dots, y_n)$ 和估计值为总体参数的估计,并简记为 $\hat{\theta}$ 。

25. 区间估计和置信区间:对于总体标志 θ ,区间估计是选用两个样本统计量 $\hat{\theta}_1(y_1, \dots, y_n)$ 和 $\hat{\theta}_2(y_1, \dots, y_n)$,满足

$$P(\hat{\theta}_1(y_1, \dots, y_n) < \theta < \hat{\theta}_2(y_1, \dots, y_n)) = 1 - \alpha$$

则区间 $(\hat{\theta}_1(y_1, \dots, y_n), \hat{\theta}_2(y_1, \dots, y_n))$ 称为总体标志 θ 在置信概率 $1 - \alpha$ 下的置信区间。置信区间包含总体标志的概率 $1 - \alpha$ 称为置信概率,或置信度。区间上限 $\hat{\theta}_2(y_1, \dots, y_n)$ 称为在置信概率 $1 - \alpha$ 下置信区间的置信上限,区间下限 $\hat{\theta}_1(y_1, \dots, y_n)$ 称为在置信概率 $1 - \alpha$ 下置信区间的置信下限。置信区间的统计解释是,对总体重复进行概率抽样,每次抽取的样本量都是 n ,利用抽取的样本构造区间 $(\hat{\theta}_1(y_1, \dots, y_n), \hat{\theta}_2(y_1, \dots, y_n))$,这样的每个区间要么包含总体标志 θ ,要么不包含总体标志 θ 。在这些区间中,包含总体标志 θ 的区间个数占所构造的区间总数的比例约为 $100(1 - \alpha)\%$,不包含总体标志 θ 的区间个数占所构造的区间总数的比例约占 $100\alpha\%$ 。

1.2 常用的总体标志和样本统计量

简单随机抽样中的符号说明:总体单元总数记为 N , N 个单元的标志值依次为 $y_1, y_2, \dots, y_N$ 。按照某个变量,把 N 个基本单元分为两类,分别为 C 类和 C' 类,记属于 C 类的单元个数为 A ,属于 C' 类的单元个数为 A' , $A + A' = N$ 。样

本单元数记为 n , n 个单元的标志值依次为 $y_1, y_2, \dots, y_n$ 。按照某个变量, 把 n 个样本单元分为两类, 分别为 C 类和 C' 类, 记属于 C 类的单元个数为 a , 属于 C' 类的单元个数为 a' , $a + a' = n$ 。

1. 总体总值, 用 Y 表示, $Y = \sum_{i=1}^N y_i$

2. 总体均值, 用 $\bar{Y}$ 表示, $\bar{Y} = Y/N$

3. 总体比例, 用 P 表示, $P = A/N$

4. 总体比率, 用 R 表示,

$$R = \frac{Y}{X} = \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i}$$

其中, 总体 N 个基本单元的成对标志值分别记为 $(y_1, x_1), \dots, (y_N, x_N)$, 比率分别记为 $r_1 = y_1/x_1, \dots, r_N = y_N/x_N$, $X = \sum_{i=1}^N x_i$ 为总体总值。

5. 总体方差, 用 S^2 表示,

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2$$

6. 样本总值, 用 y 表示, $y = \sum_{i=1}^n y_i$

7. 样本均值, 用 $\bar{y}$ 表示, $\bar{y} = y/n = (\sum_{i=1}^n y_i)/n$

8. 样本比例, 用 p 表示, $p = a/n$

9. 样本比率, 用 r 表示,

$$r = \frac{y}{x} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$$

其中, 样本中 n 个基本单元的成对标志值分别记为 $(y_1, x_1), \dots, (y_n, x_n)$, 比率分别记为 $r_1 = y_1/x_1, \dots, r_n = y_n/x_n$, $x = \sum_{i=1}^n x_i$ 为样本总值。

10. 样本方差, 用 s^2 表示,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

1.3 习题解答

1. 假设总体包含 5 个单元, 单元的标志值依次为 5.5、4.6、4.8、5.2、4.9。试计算总体总值、总体均值和总体方差。从总体中等概率无放回抽取 3 个单元作为

样本,给出所有可能样本,计算样本均值、样本方差。计算样本均值的期望和方差,以及样本方差的期望,并与总体均值和总体方差比较。

证明:由题目可知,总体单元数 $N=5$,单元的标志值分别为 $y_1=5.5, y_2=4.6, y_3=4.8, y_4=5.2, y_5=4.9$,可以计算

总体总值为

$$Y = \sum_{i=1}^N y_i \\ = 5.5 + 4.6 + 4.8 + 5.2 + 4.9 = 25.0$$

总体均值为

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i \\ = \frac{5.5 + 4.6 + 4.8 + 5.2 + 4.9}{5} = 5.0$$

总体方差为

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2 = 0.125$$

样本量 $n=3$,所有可能的样本共计 10 个,见表 1.1 的第 2 列。

样本均值的公式为

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

样本方差的公式为

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

样本均值和样本方差的计算结果见表 1.1 的第 3 和 4 列。每个样本出现的概率见表 1.1 的第 5 列。

表 1.1 习题 1 的所有样本

序号	样本	样本均值	样本方差	概率
1	(5.5, 4.6, 4.8)	4.9667	0.2233	0.1
2	(5.5, 4.6, 5.2)	5.1000	0.2100	0.1
3	(5.5, 4.6, 4.9)	5.0000	0.2100	0.1
4	(5.5, 4.8, 5.2)	5.1667	0.1233	0.1
5	(5.5, 4.8, 4.9)	5.0667	0.1433	0.1

续表

序号	样本	样本均值	样本方差	概率
6	(5.5, 5.2, 4.9)	5.2000	0.0900	0.1
7	(4.6, 4.8, 5.2)	4.8667	0.0933	0.1
8	(4.6, 4.8, 4.9)	4.7667	0.0233	0.1
9	(4.6, 5.2, 4.9)	4.9000	0.0900	0.1
10	(4.8, 5.2, 4.9)	4.9667	0.0433	0.1

样本均值的期望为

$$E(\bar{y}) = \frac{1}{10} \sum_{k=1}^{10} \bar{y}_k = 5.0000$$

每个可能样本的均值并不总是等于总体均值,样本均值的期望等于总体均值。

$$E(s^2) = \frac{1}{10} \sum_{k=1}^{10} s_k^2 = 0.1250$$

每个可能样本的方差并不总是等于总体方差,样本方差的期望等于总体方差。

样本均值的方差为

$$V(\bar{y}) = \frac{1}{10} \sum_{k=1}^{10} [\bar{y}_k - E(\bar{y})]^2 = 0.0185$$

样本均值的方差不等于总体方差,并不一定等于某个可能样本的方差。总体方差、样本方差、样本均值的方差是三个不同的概念。

R 语言代码:

```

N=5;
n=3;
yi=c(5.5,4.6,4.8,5.2,4.9);
# 总体总值为
Y=sum(yi);
# 总体均值为
Ymean=Y/N;
# 总体方差为
S2=sum((yi-Ymean)^2)/(N-1);
# 所有 10 个可能的样本序号为
yNo=array(c(1,2,3,1,2,4,1,2,5,1,3,4,1,3,5,1,4,5,2,3,4,2,3,5,2,4,
```



```

5,3,4,5),dim=c(3,10));
# 所有 10 个可能的样本
ysam=array(yi[yNo[,]],dim=c(3,10));
# 所有 10 个可能的样本均值
yimean=rep(0,10);
# 所有 10 个可能的样本方差
syi2=rep(0,10);
for(loop in c(1:10) )
{yimean[loop]=round(mean(ysam[,loop]),4);
syi2[loop]=round(var(ysam[,loop]),4)}
# 所有 10 个可能的样本均值的期望
ymean=round(mean(yimean),4);
# 所有 10 个可能的样本方差的期望
sy2=round(mean(syi2),4);
# 所有 10 个可能的样本均值的方差
ymeanvar=round(var(yimean),4);

```

2. 假设总体包含 5 个单元,其中第 1,3,4 单元属于 C 类,第 2,5 单元属于 C' 类。试计算 C 类单位总数、总体比例和总体方差。从总体中等概率无放回抽取 3 个单元作为样本,给出所有可能样本,计算样本比例、样本方差。计算样本比例的期望和方差,以及样本方差的期望,并与总体比例和总体方差比较。

证明:由题目可知,总体单元数 $N=5$ 。令属于 C 类的单元标志值为 1。则单元的标志值分别为 $y_1=1, y_2=0, y_3=1, y_4=1, y_5=0$, 可以计算

C 类单位总数为

$$\begin{aligned}
 A &= \sum_{i=1}^N y_i \\
 &= 1 + 0 + 1 + 1 + 0 = 3
 \end{aligned}$$

总体比例为

$$\begin{aligned}
 P &= \frac{A}{N} \\
 &= \frac{3}{5} = 0.6
 \end{aligned}$$