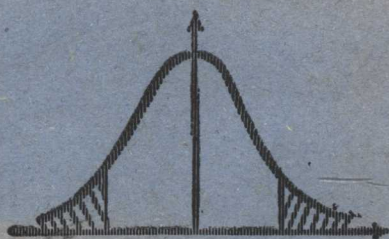


全国中等林业学校试用教材

数理统计

辽宁省林业学校 主编



林业专业用

中国林业出版社

全国中等林业学校试用教材

数 理 统 计

辽宁省林业学校 主编

林 业 专 业 用

江苏工业学院图书馆
藏书章

中国林业出版社

前 言

本教材是根据全国中等林业学校四年制教学计划和大纲的要求，在林业部教育司和林业专业教材编委会领导下进行编写的。

教材主要介绍概率论基础知识和数理统计方法的初步知识。有关的基本概念与方法，力求简单扼要，步骤清晰。在数学原理方面，受到中专数学基础的限制，做不到理论完整与严谨，有些问题只得采用说明的方法或直接给出应用的条件与结论。本教材第一章第四节一四，第四章第四节一五，第七章第四、五、六节可以根据学校的具体情况和学生的接受能力选讲。

本教材是由辽宁省林业学校于尔循、山西省林业学校王绍宗两位同志执笔的。本教材曾参考了河南农学院罗鸣福副教授为全国中等林业学校三年制编写的数理统计教材初稿。

本教材还邀请了山东省林业学校郭先平、湖北省林业学校白绍寅、广西林业学校梁祖仁、牡丹江林业学校孙永涵、洛阳林业学校单之星、南京林业学校黄清民、云南省林业学校赵思嘉、辽宁省林业学校王长利等同志参加审稿。北京林学院符伍儒副教授也提出了很好的意见。在教材编写工作中，南京林业学校、山西省林业学校等给了大力支持与协助。在此一并表示衷心的感谢。

由于我们水平所限，加以编写时间仓促，教材中难免有错误和不妥之处，请批评指正。

编 者

一九八二年十二月

目 录

前 言

绪 论	(1)
-----	-----

第一章 数据的整理与分析	(3)
--------------	-----

第一节 总体、单元、样本	(3)
--------------	-----

一、随机试验	(3)
--------	-----

二、几个基本概念	(4)
----------	-----

第二节 简单随机抽样	(5)
------------	-----

一、抽签法	(5)
-------	-----

二、随机数法	(7)
--------	-----

三、经验数据法	(8)
---------	-----

第三节 数据的整理与频率分布	(9)
----------------	-----

一、离散型随机变量的数据整理	(9)
----------------	-----

二、连续型随机变量的数据整理	(10)
----------------	------

第四节 样本特征数	(12)
-----------	------

一、样本平均数	(12)
---------	------

二、样本的方差和标准差	(14)
-------------	------

三、样本方差的计算	(16)
-----------	------

四、数据分组整理时样本特征的简捷计算	(18)
--------------------	------

五、变异系数	(19)
--------	------

第五节 总体特征数	(21)
-----------	------

一、总体平均数	(21)
---------	------

二、总体的方差和标准差	(22)
-------------	------

三、总体变异系数	(22)
----------	------

习题一	(23)
-----	------

第二章 随机事件及其概率	(25)
--------------	------

第一节 随机事件及其运算	(25)
--------------	------

一、随机事件	(25)
--------	------

二、事件间的关系	(26)
----------	------

第二节 事件的概率	(29)
-----------	------

一、概率的定义	(29)
---------	------

二、古典概率	(30)
--------	------

三、概率的基本性质	(32)
-----------	------

第三节 概率的加法公式和乘法公式	(34)
------------------	------

一、概率的加法公式	(34)
-----------	------

二、条件概率	(34)
三、概率的乘法公式	(36)
习题二	(38)
第三章 随机变量及其概率分布	(41)
第一节 随机变量	(41)
一、随机变量的概念	(41)
二、离散型随机变量的概率分布	(42)
三、二点分布	(43)
四、二项分布	(43)
五、泊松分布	(45)
六、连续型随机变量	(47)
七、分布函数	(48)
第二节 正态分布	(50)
一、正态分布的概率密度函数	(50)
二、正态概率密度曲线	(52)
三、标准正态分布概率计算	(53)
四、正态分布概率计算	(54)
五、正态分布的分位数	(56)
六、用正态分布近似计算二项分布的概率	(57)
第三节 随机变量的数学期望和方差	(59)
一、离散型随机变量的数学期望	(59)
二、连续型随机变量的数学期望	(61)
三、随机变量的方差	(62)
四、数学期望和方差的性质	(63)
第四节 χ^2 分布、 t 分布和 F 分布	(65)
一、 χ^2 分布	(65)
二、 t 分布	(68)
三、 F 分布	(69)
习题三	(70)
第四章 参数估计	(72)
第一节 统计量及其分布	(72)
一、统计量	(72)
二、样本平均数的分布规律	(72)
三、总体参数的无偏估计量	(74)
四、样本方差的分布规律	(76)
第二节 总体平均数的大样本估计方法	(76)
一、定值估计和区间估计	(76)
二、总体平均数的大样本估计方法	(77)
三、大样本单元数的确定	(81)
第三节 总体平均数的小样本估计方法	(84)

第四节 总体频率的估计方法	(86)
一、总体频率与样本频率	(86)
二、利用二项分布估计总体频率	(87)
三、利用正态分布估计总体频率	(89)
四、大样本单元数的确定	(90)
五、用泊松分布估计总体频率	(91)
习题四	(93)
第五章 假设检验	(95)
第一节 假设检验概述	(95)
一、什么是假设检验	(95)
二、假设检验解决什么问题	(95)
三、小概率原理	(95)
四、两类错误	(97)
第二节 一个正态总体平均数的假设检验	(97)
一、已知方差 σ^2 的情形	(98)
二、未知方差 σ^2 的情形	(99)
三、总体分布类型不确定或不遵从正态分布的情形	(101)
四、单侧检验	(102)
第三节 两个正态总体的平均数的假设检验	(105)
一、未知方差 σ_1^2 、 σ_2^2 的情况	(105)
二、已知方差 σ_1^2 、 σ_2^2 的情况	(108)
第四节 总体频率的假设检验	(109)
一、一个总体频率的假设检验	(109)
二、两个总体频率的差异显著性检验	(110)
习题五	(111)
第六章 方差分析	(114)
第一节 单因素方差分析	(114)
一、各组内试验次数相等的情况	(115)
二、各组内试验次数不等的情况	(122)
第二节 多重比较	(124)
一、 q 检验法	(124)
二、 S 检验	(125)
第三节 双因素方差分析	(127)
一、不考虑交互作用的情况	(128)
二、考虑交互作用的情况	(131)
习题六	(138)
第七章 试验设计与分析	(140)
第一节 试验设计的基本概念和原则	(140)
一、指标	(140)
二、因素	(140)

三、水平	(140)
四、试验设计的基本原则	(141)
五、试验的一般步骤	(142)
第二节 几种比较简单的试验设计与分析	(142)
一、完全随机化试验	(142)
二、对比排列法	(145)
第三节 随机区组试验	(145)
一、随机区组试验设计与分析	(145)
二、算题	(146)
第四节 拉丁方试验设计与分析	(150)
一、拉丁方试验设计的概述	(150)
二、算题	(151)
第五节 平衡不完全区组试验设计与分析	(154)
一、平衡不完全区组试验设计的概述	(154)
二、算例	(155)
第六节 正交试验设计与分析	(158)
一、正交试验的概述	(158)
二、正交表简介	(159)
三、用正交表设计试验的步骤	(161)
四、试验结果的计算与直观分析	(163)
五、正交试验的方差分析	(164)
六、交互作用的分析	(165)
习题七	(168)
第八章 回归分析	(173)
第一节 回归与相关的概念	(173)
一、变量之间的确定性关系	(173)
二、变量之间的不确定关系	(173)
三、总体条件平均数	(174)
四、回归方程	(175)
第二节 直线回归	(175)
一、回归直线的求法	(175)
二、计算回归方程的实例	(178)
三、相关系数及其显著性检验	(179)
四、回归问题的方差分析	(181)
五、算例	(184)
第三节 根据回归方程进行预报	(186)
一、总体条件平均数的估计值	(186)
二、用回归值及其标准差进行预报	(187)
三、用回归值预报因变量	(189)
第四节 可化为直线的曲线回归	(191)

一、曲线可化为直线的几种类型	(192)
二、算例	(195)
第五节 曲线回归	(200)
一、正则方程组的建立	(201)
二、确定回归方程	(201)
三、相关指数的计算	(202)
四、估计值的误差限及预报区间	(203)
第六节 多元线性回归	(205)
一、多元线性回归的概念	(205)
二、二元线性回归的正则方程组	(205)
三、算例	(207)
四、二元线性回归的方差分析	(209)
五、复相关系数与偏相关系数	(211)
习题八	(213)
附表1 随机数表	(217)
2 泊松 (poisson) 分布表	(219)
3 泊松 (poisson) 分布参数 λ 的置信区间表	(224)
4 正态分布的密度函数表	(225)
5 正态分布表	(226)
6 正态分布的双侧分位数 (u_α) 表	(228)
7 χ^2 分布表	(228)
8 二项分布参数 P 的置信区间表	(230)
9 t 分布表	(234)
10 t 分布的双侧分位数 (t_α) 表	(235)
11 F 检验的临界值 (F_α) 值	(236)
12 多重比较中的 q 表	(241)
13 多重比较中的 S 表	(242)
14 检验相关系数 $\rho = 0$ 的临界值 (r_α) 表	(243)
15 正交拉丁方表	(244)
16 平衡不完全区组设计表	(245)
17 正交表	(246)
18 常用的概率分布	(248)

绪 论

从小学到中学我们所学过的数学知识有一个显著的特点，就是确定性。二加一等于多少？直角三角形的勾股弦之间的关系是怎样的？一元二次方程的根有什么规律？这些问题都是确切而肯定的。自然界和社会上发生的现象中也有类似的情况，例如，向上抛一石子，石子必然要下落；高处的水，必然要向低处流动；纯水在一个标准大气压下，加热到 100°C 必然沸腾；这些在一定的条件下必然发生的现象称为确定性现象。但是在自然界和社会上的现象中，还有大量的现象呈现出不确定性。例如，抛掷一枚质地匀称的硬币，是正面朝上还是背面朝上，这在抛掷之前是无法肯定的；播种一粒种子，是发芽还是不发芽，这在播种前也是无法肯定的。这类现象归纳起来，可以看作在相同条件下一系列的试验或观察，而每次试验或观察的可能结果不止一个，在每次试验或观察之前无法预知确切的结果。这就是这类现象所呈现出的不确定性。

人们经过长期社会实践，逐步地认识到，这类现象虽然就每次试验或观察结果来说，它具有不确定性，但在大量重复试验或观察下，它的结果却呈现出某种规律性。例如，多次重复地抛掷一枚质地匀称的硬币，正面朝上与背面朝上发生的次数的比率总是近于一比一，而且抛掷的次数越多，越接近这个比率；若作相当数量的种子发芽实验，则发芽种子所占的比率，稳定于某一数值；等等。这种在大量重复试验或观察中所呈现出的固有规律性，就是我们以后所说的统计规律性。

这类现象，在个别试验中呈现出不确定性；在大量重复试验中，又具有统计规律性，我们称之为随机现象。概率论与数理统计就是研究和揭示随机现象的统计规律性的一门数学学科。

从随机现象的统计规律性出发，应用数学方法，研究随机现象各种结果发生可能性的大小，进而对所研究的问题作出估计或推断。例如，根据大量的气候观测资料，可以对未来的天气变化作出相当准确的预报；根据一定数量的样地调查资料，可以对全国或全省（县）的森林蓄积量作出令人信赖的估计；这些预报和估计的方法就是数理统计方法。

概率论与数理统计是怎样的关系呢？概括地说，概率论是数理统计的理论基础，数理统计是在概率论的基础上，着重对试验的统计资料进行整理、分析，对所研究的问题作出估计或推断。

数理统计从十九世纪末开始形成一门独立学科。当时，它的应用范围仅限于天文、测量和遗传等等方面。由于科学技术的不断发展，数理统计的应用范围不断扩大，特别是近半个世纪以来，进展尤其迅速。

我国在解放以前，数理统计的理论和应用没有受到应有的重视，长期处于停滞状态。解放初期，在中国共产党和政府的领导下，我国工农业生产以及科学技术、文化教育等事

业都有了迅速的发展，数理统计也开始在工农业生产及科学研究中得到了广泛的应用。目前，数理统计方法的应用几乎已遍及所有科学技术领域、工农业生产和国民经济的各个部门之中。例如，应用数理统计的理论与方法可以进行气象预报、水文预报及地震预报，产品的抽样验收；在研制新产品时，为寻求最佳生产方案，可用以进行试验设计和数据处理；在自动控制中可用以给出数学模型，以便通过电子计算机来控制工业生产等等。无疑，这对我国科学技术与工农业生产的发展都起到了一定的积极作用。反过来，科学技术与生产的发展又进一步促进了数理统计的理论与方法的探讨。同样，数理统计的理论与方法也在林业生产与科学研究中得到了广泛的应用，大致可分为三方面：

一、林业科学技术方面：在遗传、育种、造林、育苗、病虫害防治、化学保护等方面的科研工作中，常用到参数估计、假设检验和试验设计与分析等方法。

二、林业调查设计方面：在森林资源清查以及林区、林场的规划设计等生产活动中以及与此有关的森林学、测树学、航测、规划设计等学科领域中，常用参数估计、回归分析与抽样技术等方法。

三、森林工业方面：在森林工业产品的质量检查与质量控制等方面，常需应用参数估计、试验的设计与分析以及产品的质量检查与控制等方法。

由于数理统计的理论与方法已渗透到林业生产与科学研究的各个学科领域之中，它已成为从事林业生产与科学研究有力的工具之一。因此，中等林校林业专业、森林保护专业开设数理统计课，使学生能够初步的掌握和运用这一工具是非常必要的。应该指出的是：应用数理统计的理论与方法去从事林业生产和科学研究时，必须在林业科学理论的指导下进行，任何背离林业科学理论的研究都是徒劳无益的。

第一章 数据的整理与分析

研究随机现象的统计规律性，需要进行观测或实验，取得原始资料。利用这些资料，对所研究的问题作出估计或推断。取得原始资料的方法有两大类。第一类采用全面调查、重点调查或典型调查方法；第二类采用抽样调查方法，即由全部研究对象中抽取一部分进行调查。在林业生产和科学实验中，由于要观测的对象数量多，面积大，有些观测或实验具有一定的破坏性，所以一般多采用第二类方法。这种根据抽样调查资料，应用数学原理进行估计或推断的方法就是数理统计法。它包括两个组成部分：一是如何从全部对象中抽取部分观测、实验的对象，这是抽样方法问题；二是如何对观测、实验的结果，即对一批数据资料进行整理和分析，进而对所研究的问题作出科学的估计和推断。本章主要讨论最简单的抽样方法和数据整理，其它问题将在第三章以后逐一讨论。

第一节 总体、单元、样本

一、随机试验

为了揭示随机现象的规律性，需要进行大量的观测和实验，我们把这种观测和实验通称为试验。在林业生产中，试验也称为调查。如果这个试验在相同的条件下可以重复进行，而且每次试验结果事先不可预言，我们称它为随机试验。

例(1.1) 对同种木材，相同规格的10个试件作横纹抗压强力试验。其10次试验数据是：482, 493, 457, 471, 510, 446, 435, 418, 394, 469 (单位：kg/cm²)。

例(1.2) 从某一片同龄人工林中抽取50株林木测得胸径资料如下：(单位：cm)
11.7, 13.6, 12.0, 7.0, 12.6, 11.2, 10.2, 10.0, 14.2, 14.5, 10.8, 7.2, 5.1, 11.6, 8.5, 7.2, 4.2, 10.3, 8.7, 10.5, 11.1, 7.6, 11.3, 6.6, 9.0, 9.4, 13.0, 9.5, 10.0, 9.7, 11.5, 8.4, 9.5, 10.5, 9.3, 8.0, 11.1, 7.1, 5.9, 8.7, 12.5, 8.5, 10.2, 10.9, 9.1, 12.2, 10.6, 8.2, 10.3, 10.2。

在例(1.1)中，我们研究木材的横纹抗压强力。某一试件的抗压强力值在试验前是不能预言的，随机试验的结果以10个不同的数值表示。

在例(1.2)中，我们研究木材的胸径，抽取测定50株林木的胸径就是进行了50次随机试验。试验的结果由50个数值表示。

在中学的物理、化学等一些学科里，我们接触的往往是确定性现象。对同一研究对象不论做多少次试验，都会显示出一个确定的结果。我们只要通过一次试验就可以认识该对象的规律了。但是，随机现象就不是这样，它需要进行随机试验才能发现其统计规律性。

通过随机试验揭示随机现象的统计规律性是认识随机现象的基本方法。那么，怎样进

行随机试验？对试验结果如何进行分析？随机现象的统计规律性是怎样的？我们将在以后各章节中加以阐述。

二、几个基本概念

在例(1.2)中，为了揭示这片人工林的林木胸径的规律，需要进行实测。如果能在全林进行每木检尺，得出每株林木的胸径，当然是最全面地反映了这片人工林林木胸径的规律了。但是，一般说来，不需要也不允许去做这一类全面的试验。我们设法抽取一小部分林木进行检尺，以得到的数据资料去估计这片人工林的林木胸径。这就产生了研究对象的全体和部分的观念。为了今后讨论方便起见，我们引进几个数理统计的基本术语。

(一) 总体与单元 在指定的条件下，进行观测和研究对象的全体称为总体。构成总体的每一个基本单位称为单元。例(1.1)中，研究某种木材试件的横纹抗压强度，该种木材所有的可能试件构成总体，而每一个试件为一个单元。

例(1.2)中，欲测这片人工林林木的胸径，这片人工林为总体，每一株林木为一个单元。

将总体划分为单元时，可以采用构成总体的自然单位，也可以采用人为规定的单位。例如，欲调查某林区的森林蓄积量，可以规定以每一株林木作为单元，也可以规定以一定面积上的全部林木作为单元。譬如，以每0.04公顷面积上的全部林木作为单元。总之，总体由单元构成，单元是调查或观测的对象。

总体含有单元的个数，称为总体单元数或总体容量，用 N 表示。

例如，欲调查某林场的森林蓄积量，设该林场面积为 600ha，单元为 0.06ha 的方形林地，则总体为该林场，总体单元数：

$$N = \frac{600}{0.06} = 10000.$$

即总体包含 10000 个单元。

(二) 样本 从总体内抽取部分单元进行调查，以其成果对总体作出估计或推断，就是抽样调查。总体中抽取调查测定的这部分单元的全体称为样本。例(1.1)中，10 个被用来做实验的木材试件组成样本。

例(1.2)中，被抽中的 50 株林木组成样本。

样本含有单元的个数称为样本单元数或样本容量，用 n 表示。根据样本包含单元的多少可分为大样本和小样本。通常 $n \geq 50$ ，认为属于大样本， $n < 50$ ，认为属于小样本。

样本单元数与总体单元数的比称为抽样比，用 f 表示： $f = \frac{n}{N}$ 。如从 10000 个单元的总体中抽取 50 个单元组成样本，则抽样比

$$f = \frac{n}{N} = \frac{50}{10000} = 0.005.$$

总体中包含单元的个数可以是有限的，也可以是无限的。包含有限个单元的全体称为有限总体。包含无限个单元的全体称为无限总体。实际生产应用时，把抽样比很小的有限

总体看作是无限总体。譬如，某片林地有五千万株林木，若以每一株林木为一个单元，抽取 500 株林木组成样本，则抽样比： $f = \frac{500}{50000000} = 0.00001$ ，由于 f 很小，这个总体可以看作是无限总体。

(三) 标志和标志值 我们的目的是研究总体中单元的某个特征。如例 (1.2) 中研究的是林木的胸径。总体中单元所具有的某种特征称为标志。如木材的抗压强力，林木的胸径，树木的高度等都是一类标志。上述这类标志可以用数量表示，称为数量标志。单元在某项数量标志上的具体数值称为标志值。

例 (1.1) 中，总体的标志是试件的横纹抗压强力，10 次试验数据是 10 个标志值。这样各可能试件的横纹抗压强力值用一个可取不同数值的量表示。

例 (1.2) 中，总体的标志是林木的胸径。各株林木的胸径值用一个可取不同数值的量表示。抽中的 50 株林木的胸径值，即 50 次随机试验的结果：11.7, 13.6, ..., 10.2 是 50 个标志值，它们构成了样本的数据资料。

如果总体中单元的某种标志不是用数量表示，则称这种标志为非数量标志或品质标志。如树种、材质、病腐等都是非数量标志。对于这种标志可以设法把它们化为数量标志来研究。

从总体中抽取部分单元组成样本，对样本各单元进行调查，取得样本的数据资料，这种抽样调查的方法就是一种随机试验方法。从样本资料出发应用数学原理，对总体单元在某项标志上的标志值作出估计或推断就是我们的目的。

复 习 思 考 题

1. 随机现象和随机试验有何特点？试举例说明之。
2. 何谓总体？何谓样本？将总体、样本、单元与集合、子集、元素相对照比较，说明三者之间的关系。
3. 什么是抽样调查方法？怎样理解抽样调查方法就是一种随机试验方法？
4. 什么叫标志？什么叫标志值？取得样本各单元的标志值的目的是什么？

第二节 简单随机抽样

怎样从总体中抽取单元组成样本呢？抽样的方法很多。简单随机抽样是一种最基本的也是一种常用的抽样方法。这种抽样方法的主要特点是随机性，即总体中各单元被抽中的可能性完全相等。为保证这一点，主观意识完全排除，因此称为简单随机抽样或纯随机抽样。具体作法有以下三种：

一、抽 签 法

首先，给予总体的每一个单元一个不同的号码。为便利起见，通常按照自然数的顺

表 (1.1) 某林区40公顷森林每0.1公顷森林蓄积量数字分布图

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	13.0	15.3	11.2	15.3	20.0	10.6	10.0	14.7	11.8	16.5	—	—	1.2	—	3.5	—	1.8	—	—	2.4
2	12.4	10.6	13.6	13.0	16.5	14.1	19.4	21.2	13.6	8.8	10.0	—	1.2	6.5	8.8	—	10.0	3.0	1.2	4.7
3	17.7	16.5	13.6	12.4	17.1	10.6	8.2	17.7	14.7	16.5	11.8	8.2	4.7	0.6	8.8	1.2	3.0	—	—	2.4
4	16.5	11.2	12.4	11.8	15.3	11.8	22.4	13.6	11.8	15.9	14.1	6.5	3.5	2.4	—	3.0	3.0	5.3	5.3	3.0
5	10.0	8.2	11.8	15.3	14.7	13.0	13.0	11.2	8.8	11.8	14.7	15.3	8.8	5.3	7.1	—	—	9.4	4.7	3.0
6	22.4	24.7	21.7	23.0	13.0	25.9	27.8	10.0	14.7	17.1	20.0	17.1	11.8	14.1	8.2	5.9	7.1	0.6	—	—
7	25.3	20.0	13.5	27.0	27.7	27.1	23.0	26.0	24.2	17.7	14.1	20.0	13.5	15.3	10.6	15.3	12.4	7.1	3.0	0.6
8	21.2	27.7	26.5	21.2	20.6	17.1	28.9	25.9	18.3	24.7	19.4	27.7	18.3	16.5	8.8	10.0	11.8	13.6	5.3	7.1
9	22.4	28.3	24.7	30.0	10.0	31.8	27.7	30.6	17.7	20.0	17.7	27.1	14.1	7.1	12.4	7.1	18.8	17.1	15.9	9.4
10	10.0	14.1	26.5	27.7	30.6	16.5	25.3	26.5	27.1	15.9	23.6	18.8	30.0	16.5	14.7	24.1	11.8	15.9	8.2	12.4
11	27.7	33.0	25.3	21.8	17.7	35.3	33.3	25.3	19.1	19.3	17.7	17.7	17.7	11.8	8.8	10.6	11.8	18.8	7.7	16.5
12	22.4	21.2	15.9	22.4	14.1	18.3	28.3	18.8	14.7	18.3	20.6	18.3	13.0	8.8	5.9	13.0	14.1	11.2	0.6	9.4
13	27.1	31.8	20.0	27.1	21.8	25.3	26.0	20.0	14.7	25.9	25.3	7.7	16.5	24.2	15.3	19.4	10.6	22.4	5.9	14.1
14	27.7	27.7	20.6	23.6	23.0	23.0	29.4	16.5	29.4	21.2	25.9	15.9	9.4	12.4	21.2	10.0	15.9	12.4	21.8	20.0
15	13.0	21.8	6.5	17.1	16.5	19.4	17.1	20.6	31.2	9.4	15.3	11.8	17.1	7.1	13.6	14.7	8.8	10.0	15.3	12.4
16	21.8	13.0	11.8	13.0	8.2	17.1	14.7	12.8	17.7	18.3	15.9	9.4	12.4	21.2	10.0	15.9	12.4	10.0	8.2	7.1
17	10.6	14.7	15.3	11.8	15.9	15.3	15.3	13.0	11.2	17.7	8.8	1.2	4.1	1.8	2.4	8.8	5.3	4.1	—	1.8
18	13.0	20.0	19.4	10.0	14.1	16.5	15.3	14.7	17.7	19.4	10.6	3.5	—	1.8	—	—	3.5	3.0	4.1	3.5
19	7.7	15.5	15.9	15.9	13.3	11.8	12.4	12.4	9.4	15.9	7.1	—	10.5	1.8	0.6	0.6	—	—	—	3.0
20	18.8	18.3	17.7	13.0	9.4	15.3	4.7	18.8	11.2	11.8	1.8	1.8	—	—	—	1.2	—	3.0	5.9	1.2

序,依次给总体各单元编上1、2、3、...等号码。然后,另制一套号签,号签的数目与总体单元数相等,每个号签各与一个单元的号码相同。将号签充分搅匀,从中抽取一定数目的号签。总体单元的号码与抽中号签号码相同者,即为被抽中的单元,由这些单元的全体组成样本。这种抽签组织样本的方法,在总体单元数不多时适用。总体单元较多时不大方便。

二、随机数法

随机数法是利用随机数表组织样本的方法。随机数表是用抽签的方式抽出大量数字制成的表。因此,用随机数表抽取样本,不论抽出哪个数字或哪一段数字确定样本都是随机的,在本质上与抽签法是相同的。但是用这种方法确定样本,不论总体单元多寡都很方便。

目前,已编印出版的随机数表有许多种。附表1是中国科学院数学研究所概率统计室编印的《常用数理统计表》中的随机数表。该表只有两页,每页有50行50列共2500个数字。现以例(1.3)说明应用随机数表组织样本的方法。

例(1.3) 以某林区中40ha林地上的林木为总体,以每0.1ha矩形林地上的林木为单元,总体中400个单元的蓄积量如表(1.1)所示。试按简单随机抽样方法抽取20个单元组成样本。

按简单随机抽样方法,应用随机数表组织样本具体作法如下:

第一步 按顺序将总体各单元编号。可以有形编号(直接编上号码),也可以无形编号(不一定直接编上号码,但有一定方法知道这个号码对应的单元在什么地方)。号码的顺序一般从左到右,由上到下排列。本例每行20个号码,共20列400个号码。

第二步 确定样本单元数 n ,本例 $n=20$ 。

第三步 按随机数表确定抽中单元的号码。先确定起点的页数、行数和列数。用铅笔在随机数表上随机指定一点。落点后的数字或距落点最近的数字为奇数,则起点为第一页,若为偶数,则起点为第二页。然后另行指定一点,根据落点的两位数字确定起点的行数。由于每页只有50行,所以,如落点的行数大于50,则将落点的两位数减去50。如果落点的两位数是00,则加50。确定了起点的行数以后,再用同样的方法,另指定一点确定起点的列数。

确定了起点之后,就可确定所抽中单元的号码。从起点开始向右或向下取足所抽取的号码。通常,如果所抽取的是一位数或两位数,则依次向右取,到达该行右端,便移到下一行左端,继续向右取。如果所需抽取的数字是三位或三位以上的数字,则从起点开始,依次向下方取,到达下端后,按顺序转到相邻列的最上行,然后从该行开始,继续由上向下取,直到取足所需数字为止。

如本例,用上述方法确定起点为第一页第17行第10列。因为总体有400个单元,编号为三位数,所以,从起点向右取三个数字,依次向下取,凡是小于400的数字都记下来,直到取足20个数字为止。现在,第一个数字是596大于400剔除,由这个数字向下看,依次取得的数字是:233,350,55,348,166,33,175,33,97,22,309,358,171,94,

373, 39, 245, 171, 160, 395, 这20个数字则为被抽中单元的号码。

如果被抽取的各单元不允许重复, 则抽中的单元号码是: 233, 350, 55, 348, 166, 33, 175, 97, 22, 309, 358, 171, 94, 373, 39, 245, 160, 395, 332, 271。

第四步 根据抽中的号码, 到总体中找到相对应的单元。然后收集所需要的单元标志值, 得到样本数据资料。

本例中, 若抽中单元允许重复, 称为重复抽样, 则样本数据资料如表 (1.2); 若抽中单元不允许重复, 称为非重复抽样, 则样本数据资料如表 (1.3)。

表 (1.2) 重复抽样样本数据

抽中号码	233	350	55	348	166	33	175	33	97	22
标志值	13.0	19.4	8.8	14.7	31.8	1.2	12.4	1.2	0	10.6
抽中号码	309	358	171	94	373	39	245	171	160	395
标志值	17.7	3.0	17.7	5.3	10.5	1.2	21.8	17.7	7.1	0

表 (1.3) 非重复抽样样本数据

抽中号码	233	350	55	348	166	33	175	97	22	309
标志值	13.0	19.4	8.8	14.7	31.8	1.2	12.4	0	10.6	17.7
抽中号码	358	171	94	373	39	245	160	395	332	271
标志值	3.0	17.7	5.3	10.5	1.2	21.8	7.1	0	1.2	25.9

三、经验数据法

当无限总体以及有限总体单元个数非常多时, 如何进行简单随机抽样? 此时, 抽签法与随机数法都不适用。在这种情况下, 可以把满足试验条件所进行的每一次试验作为一个样本单元, 只要按照所规定的样本单元数 n , 以所进行的 n 次试验所得到的数值作为样本数据资料就可以了。这种组织样本的方法称为经验数据法。譬如: 欲了解堆放在一起的10000斤小粒种子的发芽率, 需抽取1000粒种子作发芽试验, 此时看作是无限总体的简单随机抽样问题。只要不凭主观意志随便取出1000粒种子进行发芽试验就可以了。

从总体中抽取单元组织样本的方法很多, 除了简单随机抽样以外, 常用的抽样方法还有分层抽样, 系统抽样, 整群抽样, 二阶抽样等等。这些抽样方法将在测树学中讲述。

复习思考题

1. 简单随机抽样有什么特点?
2. 班上有五张电影票, 谁去看电影呢? 采取抓纸团的方法, 如何组织才符合简单随机抽样原理?

第三节 数据的整理与频率分布

例 (1.1) 中, 木材的各可能试件的抗压强度和例 (1.2) 中, 总体各株林木的胸径都用可取不同数值的量表示。如果从总体中随机抽取一个单元, 抽中单元的标志值, 也将由这些量表示。因而这些量可以表示随机试验的结果, 通常称为随机变量。从抽样调查的角度出发, 我们将总体单元的标志值看作随机变量, 而某个单元的标志值看作是随机变量取得的一个数值。譬如例 (1.2) 中, 林木的胸径值是随机变量, 如抽出的第一株林木的胸径值 11.7 就是这个随机变量取得的一个数值。这样, 在抽样调查中, 样本的数据资料就是某个随机变量可能取得的一些数值。放在我们面前的样本数据往往是一堆散乱的数字, 为了研究随机变量的特征, 首先需要一番加工整理, 这就是数据整理。

一、离散型随机变量的数据整理

例 (1.4) 为了调查某林场落叶松造林成活率, 随机抽取 50 行, 每行随机调查 50 穴。经调查各行成活的穴数如下:

42, 41, 39, 44, 43, 38, 47, 45, 42, 34, 38, 43, 40, 44, 46, 40, 41, 43, 42, 46, 42, 41, 41, 43, 37, 40, 43, 45, 49, 42, 36, 41, 42, 43, 44, 44, 48, 45, 39, 42, 41, 43, 40, 44, 42, 38, 40, 42, 48, 43.

在这个问题中, 总体是该林场的落叶松造林地, 以 50 穴为一个单元, 抽取 50 个单元组成样本。总体单元的标志值即每 50 穴中成活的穴数是一个随机变量。从理论上说, 这个随机变量仅可以取 0, 1, 2, ..., 50 这 51 个数。象这类只能取有限个或无穷可数多个数值的随机变量称为离散型随机变量。

从总体中抽取了 50 个单元组成样本, 我们看到样本单元的标志值中有许多是同一数值, 所出现的不同数值只有 15 个, 它们出现的次数是不同的。统计这些数值出现的次数整理成表 (1.4), 称为样本频率分布表。

表 (1.4) 落叶松造林成活穴数频率分布表

成活穴数 x_i	34	36	37	38	39	40	41	42	43	44	45	46	47	48	49	总和
频数 f_i	1	1	1	3	2	5	6	9	8	5	3	2	1	2	1	50
频率 $\frac{f_i}{n}$	0.02	0.02	0.02	0.06	0.04	0.10	0.12	0.18	0.16	0.10	0.06	0.04	0.02	0.04	0.02	

在表 (1.4) 中出现了两个名称, 频数和频率。数值 x_i 出现的次数称为 x_i 的频数, 记作 f_i 。不难看出, 各个 x_i 的频数的总和, 即总频数等于样本单元数 n : $\sum_{i=1}^n f_i = n$ 。 x_i 出现的频数与总频数的比 $\frac{f_i}{n}$ 称为数值 x_i 发生的频率, 因为 $0 \leq f_i \leq n$, 所以 $0 \leq \frac{f_i}{n} \leq 1$ 。即任何数值 x_i 发生的频率在 0 和 1 之间, 并且总频率等于 1。