数学基础知识丛书

数理统计初步

胡宣达

江苏人民出版社

数理统计初步

胡宣达

江苏人民出版社

数理统计初步

胡宜达

*

江苏人民出版社出版
江苏省新华书店发行
江苏新华印刷厂印刷

1980年8月第1版　1980年8月第1次印刷
印数：1—17,000册

书号：13100·059　定价：0.57元

责任编辑　赵遂之　何震邦

内 容 提 要

这套《丛书》共二十四册，系统介绍数学基础知识和基本技能，供中学数学教师、中学生以及知识青年、青年工人阅读。

《丛书》根据现行全日制十年制学校《中学数学教学大纲》（试行草案）精神编写，内容上作了拓宽、加深和提高。《丛书》阐述的数学概念、规律，力求符合唯物辩证法，渗透现代的数学观点和方法，以适应四个现代化的需要。为了便于读者阅读，文字叙述比较详细，内容由浅入深，由易到难，循序渐进，习题、总复习题附有答案或必要的提示。

本书共分七部分，第一、二两部分介绍数理统计的目的、步骤和一些最基本的概念，第三部分讲概率的概念和算法，第四、五部分讲随机变量的一些性质，第六部分讲统计检验，第七部分讲回归分析。

本书在编写过程中，曾得到俞中明同志的宝贵帮助。

目　录

一、引　论

概率论与数理统计是数学的一个重要领域。它的方法已被广泛地应用于各门自然科学及国民经济各部门，因此学习它的基本理论和方法也就具有重要的意义。

本书初步介绍了这一学科的基础知识。为了使概率论中的一些概念和理论更好地联系实际；同时也为了使读者在阅读上能由浅入深，循序渐进，我们把概率论与数理统计作为一个统一的数学题材来处理，并以数理统计作为全书前后联系的一条线索，贯串始终，这是本书所以定名为《数理统计初步》的原因，然而本书的重点与难点却在概率论部分。

本书的一至三章的内容，一般只要具有高中文化程度的读者，就可以较容易地阅读。从第四章开始，我们是假定读者已具有微积分初步知识来编写的。关于微积分的初步知识可参考本丛书的《微积分初步》。至于本书所涉及的多元函数微积分的知识，读者可参考樊映川等编的《高等数学讲义》下册。对于没有具备这方面基本知识的读者来说，可暂时略去那些理论证明与公式推导。

§1　数理统计研究的对象

在自然界中广泛存在一类随机现象（即不确定的现象）。例如，在日常生活中，我们观察某一个公共汽车站上等车的人数，就会发现，有时很空，有时很拥挤，要排长队，从而表现出某种偶然的性质，这就是一种随机现象。又如，我们在进行某

种测量时，由于种种偶然因素的影响（如测量仪器受大气的影响，观察者生理上或心理上的变化等）不可避免地会产生测量误差。这种误差也就是我们在一般数据处理中的所谓偶然误差或随机误差，它也是一种随机现象。再如在容器里盛着一定体积的气体，气体分子由于受到其它分子的冲击而产生运动速度和方向的随机变化；飞机在高空飞行时由于大气中湍流等各种影响，它环绕着重心作随机摆动；船舶在海洋中航行时由于受到海洋波浪的影响而产生各种各样的摆动（纵摆、横摆以及高低起伏等等），这些都是随机现象。

自然界中客观存在的随机现象，尽管表现形式各不相同，但它们都具有某种偶然性的共同特征。正如恩格斯在《路德维希·费尔巴哈和德国古典哲学的终结》一书中所指出的："被断定为必然的东西，是由纯粹的偶然性所构成的，而所谓偶然的东西，是一种有必然性隐藏在里面的形式"（人民出版社，1972年4月第1版，第35页）。科学的任务就在于从各种偶然性中发现潜在的必然性即规律性，进而利用这种规律性来达到改造自然的目的。实践证明，当我们研究了大量的同类随机现象后，通常总会揭露出一种完全确定的规律性。例如，在观察某公共汽车站上等车的人数的例子中，假如我们经常地留意观察，时间久了我们自然就会找出它的规律性。在一天中哪段时间较空，哪段时间较挤，什么时候是其高峰，等等。又如，在打靶中，当射击次数不大时，靶上命中点的分布是杂乱无章的，没有什么显著的规律性。当射击次数增加时，分布就开始呈现一些规律性。射击次数越大，规律性越清楚。再如，在容器里盛着一定体积的气体，从分子物理学的观点来看，气体是由无数气体分子所组成的，这些分子在不断运动着，且在运动过程中互相影响着，因而每个分子的运动轨道、速度、

方向都是随机的。但是我们知道，从宏观来看，气体对容器壁的压力却是稳定的，这是因为分子数目足够大，因而各个分子的运动所具有的随机性在集体作用下就相互抵消了。诸如此类，这种规律性是大量随机现象所特有的一种规律性，也就是所谓"统计"规律性。通过上面这些例子，我们可以清楚地看到，统计规律性具有规律的一切特征：普遍性、客观性、必然性。然而统计规律性又不同于一般的动态规律性。统计规律性，首先它只适用于同类随机现象的整体。例如，上面谈到的气体对容器壁的压力，具有稳定的数值，这是一种统计规律性，它只适用于这个容器的气体分子的整体，认识了这个统计规律性并不能预言容器内每个气体分子运动的确切状态。其次，统计规律性也只有通过对同类随机现象进行大量的观察才能被发现。

概率论与数理统计就是研究大量随机现象的这种统计规律性的科学，但数理统计与概率论又有何不同呢？大体说来，概率论着重对客观的随机现象提出各种不同的理想化了的数学模型并研究其内在的性质与相互联系。数理统计是以概率论为基础，着重于对统计资料进行分析、研究、验证它是否符合某种数学模型，从而作出有用的推断。前面我们已讲过，这种统计规律性要通过对随机现象进行大量的观察才能被发现，但客观上又只允许我们对随机现象进行次数不多的观察，从表面看来这是矛盾的！但是只要我们能充分利用这些观测资料和局部与整体之间的内在联系来进行分析与推断，则我们仍然能够认识这种规律性。因此，**数理统计的中心任务就是从局部观测资料的统计特性来推断事物整体的统计特性。**

因为这种从局部观测去推断整体的方法具有普遍的意义，所以数理统计的应用非常广泛。它可以应用到各门科学及

各种工业技术中去。例如在工业生产中的产品质量控制与抽样检查，气象学中的天气预报，地质勘探，地震预报，工程设计中安全系数的统计分析，农业生产中的病虫害防治，良种的选择以及国民经济的其它部门中，数理统计都有广泛的应用。随着四个现代化的迅速进展，对科学技术的要求愈来愈高，因而数理统计的应用范围也将愈来愈广泛。另一方面，由于现代工业技术和各门科学的领域都在不断地扩展着，因而新的数理统计课题也随时涌现出来要求解决，这样无疑地也将大大推动数理统计理论本身的迅猛发展。

§2 总体、个体与样本

为了今后叙述的方便，我们先引进几个最基本的概念。

1. 总体（或称母体）和个体。

我们在某一次统计分析工作中**所要研究的对象的全体称为总体（或称母体），其中的一个单元则称为个体。**例如我们要了解某砖瓦厂某一天所产青砖的抗压强度情况，那么这家砖瓦厂这一天所生产的所有青砖的抗压强度便构成我们研究对象的全体，也就是构成我们研究的总体。而该厂这一天生产的每一块青砖的抗压强度则为我们研究的一个个体。可是，如果我们现在要研究该厂最近三个月来每天所产青砖的平均抗压强度的逐日变化情况，那么这家砖瓦厂最近三个月即九十天中每天所产青砖的平均抗压强度的全体便成为研究的对象，也就是构成我们研究的总体。而某一天所产青砖的平均抗压强度则为我们研究的一个个体。

从上面所举的例子可以看出，什么是总体，什么是个体，并不是一成不变的。而要看每一次研究的任务而定，当研究的对象改变时，总体和个体也随之改变。

有时我们所研究的问题，不象青砖的抗压强度那么容易捉摸，而是比较抽象的东西。例如要研究一个蒸馏车间里的温度，我们可以用一只温度计在车间里到处去测量温度。这种测量和青砖的抗压强度很不相同。将一块青砖在强力试验机上量得其抗压强度时，这块青砖已被压碎，不可能再在同一块砖上第二次测量抗压强度，所以某一天所产的砖，如果总共是十万块，那么抗压强度至多只能有十万个。反之，在一个蒸馏车间测量温度时，因为可在任意地方测量，所以可测得数字的个数是没有限制的，从这一角度来看，青砖的抗压强度和蒸馏车间的温度这样两个总体是有所不同的。前者所包含的个体是有限多个，后者则为无限多个。数理统计中常称前者为**有限总体**，后者为**无限总体**。值得注意的是十万块青砖的抗压强度构成的总体虽为有限总体，但从统计观点来看，它已接近无限总体。

2. 样本（或称子样）

总体的性质由其中各个个体的性质而定，所以要了解总体的性质，就必须测定各个个体的性质。很容易理解，要对一个总体的性质了解得很清楚，必须把总体之中每一个个体的性质都加以测定。但把总体中所有个体都一一加以测定，在实际中常常是不可能的。第一，在很多情形下，总体中个体数目甚多，甚至近似无限多，事实上不可能把总体中所有个体都加以测定。例如一家砖瓦厂每天所产的砖。一家机器零件厂每天加工的螺钉等，就属于这种情形。第二，也有不少情形，总体中个体数目虽然并不很多，但对各个体的某种数量性质的测定是一种具有破坏性的测定（一种测定或试验称为是破坏性的，是指产品一经测定或试验，便被破坏而不能使用，例如，试验一条皮带的抗拉强度，当抗拉强度被测定时，皮带已被拉断。

再如试验灯泡的使用寿命，炮弹的爆炸力等等，都是破坏性的)，在这种情况中，总体中个体数目虽然不多，但仍不能对所有个体一一测定。

由于以上两种原因，在统计分析工作中，常从总体中任意抽取一部分个体，对这些个体来加以测定，然后根据它们来推测总体的性质。因此我们把抽取出来的**总体的一部分称为样本(或子样)。样本中所含个体的数目称为样本的容量(或样本的大小)**。正如§1中所述，数理统计的中心任务就是通过样本来了解总体。容易理解，既然要根据样本的性质来推断总体的性质，那么怎样抽取样本，样本中应包含多少个体，才能使样本的性质代表总体的性质，便成为数理统计中一个重要的问题。

最后需要强调指出的是数理统计中的一个个体和生产实际中的一个实物单元，如一块青砖，一发炮弹，一只灯泡等等是不同的概念，必须分辨清楚。因为我们研究的对象并不是这些实物单元的本身，例如它的化学成分与物理性质等，而是研究它的某一项数量特征，例如，抗压强度，爆炸力，使用寿命等等。

§3 数理统计工作的步骤

前面我们已经提到，数理统计的中心任务就是从局部观测资料的统计特性来推断事物整体的统计特性，确切地说，也就是从样本的统计特性来推断总体的统计特性。为了解决这个问题，一般需要进行以下三步：

第一步：数据的收集和整理工作。

进行这一步工作需要深入现场收集资料（这些资料必须保证是完全可靠的实际观测或试验数据，如发现有可疑数据

就必须对这些可疑数据重新测定或作必要的试验），将收集到的资料加以整理和归纳，用列表和作图的方法，并借助于少数几个简单的特征数字，把这些数据的主要特点表现出来。在数理统计工作中，所收集的资料大都为一个或几个样本的资料。例如某地震台站为了寻找地震前兆，观测收集了最近五天内，该地区形变电阻率 $\rho_K(\Omega m)$ 的一批数据资料，数理统计工作者，就要将这批数据整理分组，计算平均数及其它有关的特征数，使有关形变电阻率的主要情况能够突出地表现出来。这就是数理统计工作的第一步。

第二步：分析性工作。

这一步工作的内容是将整理、归纳后所得到的数据资料加以分析，发掘这些数据资料所遵循的规律。

第三步：推断性工作。

通常收集来加以整理和分析的统计资料，绝大多数是一个或几个样本的资料，已如上述，当我们分析这些资料，并在各门科学理论的指导下，发现它们所遵循的规律以后，就要根据所发现的规律来对提供这一个或几个样本的总体所遵循的规律进行推测性的判断。有时还需要根据这样推断出来的规律，去预测事物未来发生的情况。例如我们收集了某一条河流过去数十年或百余年中，每年七月份上半月和下半月河水流量的统计资料，经过整理和分析后，在水文学理论指引下，发现它们之间有一定的关系。例如七月上半月流量大时，下半月亦较大。但数十年或百余年的七月份流量资料，是这条河流过去、现在和未来无数年代的七月份流量资料中的一小部分，也就是说过去数十年或百余年的流量资料为一样本，而过去、现在和未来无数年代的七月份的流量则为总体。当我们从一个样本的资料中发现了每年七月份上半月和下半月流量之间的

关系后，以后每年逢到七月十五日，便可根据过去十五天的流量来预测今后十五天的流量。这便是数理统计工作的第三步，统称为推断性工作，意思就是对总体作推测性的判断。这一步工作是最有实际用处的工作。但这里应特别注意的是根据样本所遵循的规律来预测总体未来的情况时，必须样本与总体的基本条件没有改变，否则是没有任何意义的。例如上述河道，如果今年被疏浚加宽了，或被筑了拦河坝或人工湖，那么根据样本所发现的规律，就不能再用来预测未来的流量的情况了。

二、数据整理

由实验、观测或其它方法所收集到的一批资料中的各个数据通常是分散着的，好象没有系统和次序，它们所遵循的规律往往并不能一目了然，因此，我们必须去粗取精，去伪存真，对数据作科学的整理和归纳，方能显露出这一批数据所遵循的规律来。这正是上一章所述的数理统计工作的第一步。本章介绍有关数据的整理方法，使我们了解数据为什么需要整理和如何整理。

§4 几个重要的统计特征数

我们从总体任意抽出一个样本，得到了一批数据 x_1，x_2，…，x_n。为了把这批数据的主要特点表现出来，我们首先要计算几个简单的特征数字。

数理统计中常用的特征数可以分为二类。一类是表示数据的集中位置的，如样本均值，中位数等。一类是表示数据的离散性质或离散程度的，如极差，样本方差等。这几种特征数不仅是常用的，而且是基本的。但有时这两种基本的特征数还不足以说明数据的真实面貌，而需要把某两种基本特征数联合在一起使用借以说明问题。变异系数便是属于这一类复合型特征数。

下面我们就来一一地介绍这几个重要的统计特征数。

1. 样本均值

均值是表示数据集中位置的各种特征数中最基本的一

种。通常在数据处理中所用的均值，指的是算术平均值。数理统计中称这个算术平均值为**样本均值**。并记为 $\bar{x}$，即

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i. \qquad (2.1)$$

式中记号“$\sum$”是求和的意思，$\sum_{i=1}^{n}x_i$ 表示从 x_1 加到 x_n。

大家知道，观测或试验数据往往带有各种各样的误差，如观测所造成的人为误差，仪器本身的误差，大气影响所引起的误差等等。这些误差有正有负，因此求均值后，正负误差消去了一部分，从而显示了观测或试验数据的真实面貌。所以均值的作用就是消除数据中的一些局部的、随机的波动，表征数据的集中位置。

2. 中位数

我们把得到的数据 $x_1, x_2, \cdots\cdots, x_n$ 按大小次序排列，排在正中间的那一个数，数理统计中称为**中位数**。并记为 M_e。

当 n 为奇数时，正中间的数只有一个；当 n 为偶数时，正中位置的数有两个。在后一种情形，中位数等于这两个数的算术平均值。

中位数也是表示数据集中位置的一种特征数。当然，以中位数来表示数据的集中位置是比较粗略的，但它的优点是可以减少计算量。

3. 极差

极差是表示数据离散程度（或波动大小）的各种特征数中的最简单的一种。

极差是指数据中最大的一个与最小的一个之差。即极差

$$R=Max\{x_1,x_2,\cdots,x_n\}-Min\{x_1,x_2,\cdots,x_n\}.$$

式中 $Max\{x_1, x_2, \cdots, x_n\}$ 和 $Min\{x_1, x_2, \cdots, x_n\}$ 分别表示 $x_1, x_2, \cdots\cdots, x_n$ 中最大的和最小的。

用极差表示数据的离散程度(或波动大小)的意义是明显的，但由于极差只用了最大与最小的数据，没有充分利用数据提供的信息，因此反映实际情况的精度较差。于是人们想出了另一种表示数据离散程度的特征数。

4. 样本方差

样本方差的公式为

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad (2.2)$$

也可用它的平方根 S——称为**样本均方差或标准离差**，来衡量数据的离散程度。S 越大，波动越大；S 越小，波动越小。S 或 S^2 比极差 R 反映问题精确，但计算比极差要复杂，所以各有优缺点，看具体情况加以运用。

为什么要以 S^2 或 S 来衡量数据的离散程度呢？首先我们注意，求和号中的每一项 $(x_i - \bar{x})$ 是表示第 i 个数据与样本均值 $\bar{x}$ (它代表这批数据的集中位置)的离差。如果将这些离差单纯地相加，容易证明其和为零，因而无法用来表征数据的离散程度。如用这些离差的绝对值相加来表征数据的离散程度，当然可以避免上面的缺点，但给计算带来麻烦(关于这一点，读者以后将会知道)，因此我们转而采用离差的平方和来衡量。其次，大家可能要奇怪，为什么在计算样本方差时要用 $n-1$ 作为除数而不象计算样本均值那样用 n 作为除数呢？关于这一点我们将在第五章 §19 中加以说明。

5. 变异系数

变异系数是复合型特征数的一种。我们都有体会，测量较