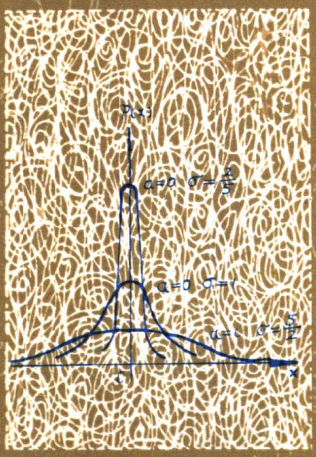

潘曾廷 编

概率论与数理统计初步

概率论与数理统计初步

潘曾挺　编

黑龙江人民出版社

1982年·哈尔滨

责任编辑：田兆民
封面设计：蒋　明

概率论与数理统计初步

潘曾挺　编

黑龙江人民出版社出版
（哈尔滨市道里森林街42号）
黑龙江新华印刷厂印刷　黑龙江省新华书店发行
开本787×1092毫米 1/32·印张 5 2/16·字数100,000
1982年11月第1版　1982年11月第1次印刷
印数1—7,400

统一书号：13093·56　　定价：0.45元

出 版 说 明

为加速实现四个现代化，迅速培养和造就大批又红又专的建设人材的需要，我们将陆续出版一套《中学生课外读物》。

这套读物包括数学、物理、化学、语文、历史、地理等基础知识和典型题解答几十种。这本《概率论与数理统计初步》就是其中的一种。

本书以全日制十年制学校高中数学的有关内容为基础，适当扩大了知识范围，全面系统、深入浅出地讲述了概率论与数理统计的初步知识。

本书可供中学生、知识青年自学之用，也可供中学数学教师参考。

中学生课外读物

统一书号：
定　　价：

目　录

一　概率统计的研究对象

自然界的各种现象和过程，就其发生和发展的规律来说，可以分为确定性现象和随机现象两大类．对于前一类现象来说，是我们比较熟悉的现象．例如求方程$ax^2+bx+c=0$的解；求函数$f(x)=ax^2+bx+c$的极值；物体的自由落体的运动规律和斜抛运动规律；物质的化合与分解；……都是确定性现象．这些现象的特点是：只要一组基本的条件确定下来之后，那么这个现象的发展变化就按照事先能够预计到的结果出现．而在有些现象中，却表现出相反的特点．例如：

1. 掷一枚质量均匀的硬币，每次投掷的结果可能出现国徽图案也可能出现币值．在投掷之前无法确定将出现哪一个结果．如果把每一次投掷作为一次试验；那么每一次试验的结果是随机的，或者说是偶然的．

2. 一位工人师傅用同一种钢材，在同一台车床上按同样的操作规程，加工口径为10毫米的螺栓．而实际加工出来的螺栓口径一般不等于10毫米，有的大于10毫米，有的小于10毫米，在每次加工之前，无法确切地预料加工后的螺栓的口径多大．如果把加工一个螺栓作为一次试验，那么每次试验的结果是随机的．

3. 有一批产品，已知它的废品率为P．今从这批产品中任意抽取一件产品，那么，这件被抽到的产品可能是废品也可能是合格品，在抽取之前无法确切预料抽到什么样的产

品。如果把每次抽取一件产品看作一次试验，那么，每次试验的结果是随机的。

上述例子所给出的现象都是随机现象，其特点是在一组基本条件确定下来之后，这个现象将出现什么结果是无法预计的，也就是说对这个现象进行试验，其试验结果是随机的。

虽然对于随机现象来说，每次试验的结果具有偶然性，但是只要对随机现象进行多次重复的试验，那么每个结果的出现都有一定的规律。我们仍以上述三例来说明：

1. 掷一枚质量均匀的硬币，每次投掷都可能出现两个结果，这两个结果中哪个都可能出现，似乎没有规律可言，但只要投掷的次数很多，就会发现必然的规律性：出现国徽图案朝上和出现币值朝上的次数大致相等。

2. 每次加工螺栓的口径不完全一样，有时大，有时小，看不出什么规律。但是只要我们对该工人师傅加工的大量螺栓进行测量之后，就会发现螺栓口径的大小是服从某种确定的规律的。例如，这批螺栓口径的平均数等于10毫米，口径大于10毫米和小于10毫米的个数大致相同；误差较小的占多数；误差较大的占少数；……。观察的螺栓越多这种规律性就越明显。

3. 每次抽查产品的结果，可能是合格品也可能是废品，好象没有规律，但是当我们重复进行大量的抽查时就会发现：如果抽取产品的总次数是 n，其中抽取到废品的次数是 v，那么抽取到废品的频率是 $\frac{v}{n}$。随着抽取总次数逐渐增加，频率 $\frac{v}{n}$ 逐渐稳定于这批产品的废品率 p 的附近。

随机现象在生产斗争和科学试验的各个领域中是普遍存

在的，不管各个领域的具体内容多么不同，所有的随机现象都具有偶然性和必然性的两重性，必然性是主导方面，是本质，也就是说，对任何随机现象进行一次观察其结果的不确定性和大量重复观察时呈现出的规律性是一对矛盾，是矛盾的对立统一。

在大量重复试验或观察中反映出来的随机现象所具有的规律性叫做随机现象的统计规律性，概率与数理统计就是研究随机现象的统计规律性的数学学科。

近年来，概率论和数理统计的发展使它实际上进入了人类活动的各个领域。它不再被认为仅仅是对数据加以整理而得到的一堆图和表，而被认为是围绕着对具有偶然性的问题作决策的一门完整科学。无论在试验一种新药品、检验产品质量、天气预报、公用服务事业、地震预报、最优方案的设计、尖端技术的研究……的时候都会碰到，所以概率论与数理统计适用于广大范畴。如果说概率论和数理统计在它发展的现阶段能处理所有包含偶然性的问题，也许有点过分。但是由于新方法在不断发展，现代概率论和数理统计至少已经为有条理地、系统地研究这些问题提供了一个数学模型，正如微积分的体系为研究确定性现象的问题提供了数学模型一样。概率论和数理统计已经成为解决各类有关随机现象的实际问题的强有力的数学工具。

二 数据整理

（一）数据为什么要整理

概率论和数理统计是从数量方面研究大量随机现象统计规律性的数学方法．我们在运用这个方法去解决实际问题时，就要对同一随机现象进行大量试验（或观察），从而得到一批试验（或观察）数据．这些数据提供了很有用的情报，可以帮助人们发现存在的问题，认识事物的内在规律．

但是这些数据提供的情报往往不是一目了然的，而是蕴藏在大量数据之中．我们必须去粗取精，去伪存真，对数据做科学的整理和分析，尽可能充分而且正确地从中提取出情报来．

（二）基 本 概 念

为了今后叙述方便，我们先介绍两个基本概念．

1. 总体与个体

总体也叫母体，它是指在一次统计分析中我们所研究的对象的全体．研究对象中的每一个单位叫做个体．

举例来说，如果我们要研究一批机器零件的毛坯的重量，那么所有这批毛坯重量（单位：公斤）便是我们研究的全部对象，因此，这些毛坯的重量的全体就是我们要研究的总体．而每一个毛坯的重量便是一个个体．如果我们要研究这批毛

坯生产中每天的毛坯重量的变化情况，那么生产这批毛坯时，所有每天生产的平均毛坯重量的全体便是我们研究的总体．每一天生产的毛坯的平均重量便是个体．可见这里所讲的一个个体与工业产品的单位是不同的概念，如何确定总体和个体，只有根据我们所要研究的具体问题来确定．随着所研究的问题的改变，研究对象就改变，这时总体和个体的含意也随之而改变．

2. 样本

总体的性质由其中各个体的性质而定，要了解总体的性质就必须测定每个个体的性质，但有时总体中个体数量太大，不可能在给定的时间内逐个测定．有些试验中个体性质的测定是破坏性的，如研究炮弹的杀伤半径，测定一个就要爆炸一个，这样要研究总体的性质，只有从总体中抽出一部分个体进行考察，这些被抽出的个体的全体叫做样本或子样，所以样本就是总体的一部分，样本中所含个体的个数叫做样本的容量或大小．

数理统计的主要内容就是要研究如何合理地抽取样本，并从对样本的分析研究中得到总体的有关性质．

（三）数据的统计表和统计图

怎样进行数据的整理呢？通常是要进行两项工作．其一，是对所抽取样本的数据进行制表和制图，以便对样本的数据的状况和隐含的规律有一个初步的直观了解．其二，根据已有的统计表和统计图，运用数学的方法计算出一些简单的统计特征数，作为研究总体性质的出发点．

由实验得到的数据可以分为离散型和连续型两类，所谓离散型的数据是指只取自然数、整数或有理数等数值的数据．如机器零件中废品的个数可以是零个，一个，二个，等等．所谓连续型数据是指可以取某一区间内的任何实数值的数据．如机器零件的长度，棉布的强力，等等．

数据列表和制图的方法可以分为不分组和分组两种．在实践中主要是对连续型数据用分组的方法列表和制图．因为数据分组后，可以使该试验数据的统计特征更直观地显示出来．对于其他类型的数据的制表和制图的方法就略去了．

分组数据的列表和制图主要包括以下内容：将数据分组；列出频数分布表；列出频率分布表；绘制频数分布图和频率分布图．具体方法通过下面的例子来说明．

在20天内从某维尼纶厂正常生产时生产报表上看到维尼纶纤度（表示纤维粗细程度的一个量）的情况，有如下的100个数据：

1.36	1.49	1.43	1.41	1.37	1.40	1.30	1.42	1.47
1.39	1.41	1.36	1.40	1.34	1.42	1.45	1.35	1.42
1.42	1.39	1.44	1.46	1.39	1.42	1.42	1.30	1.34
1.42	1.37	1.36	1.37	1.42	1.37	1.37	1.44	1.45
1.32	1.48	1.40	1.45	1.39	1.34	1.39	1.53	1.36
1.48	1.40	1.39	1.38	1.40	1.36	1.45	1.50	1.43
1.38	1.43	1.41	1.48	1.39	1.45	1.37	1.37	1.39
1.45	1.31	1.41	1.44	1.44	1.42	1.47	1.35	1.36
1.39	1.40	1.38	1.35	1.42	1.43	1.42	1.42	1.42
1.40	1.41	1.37	1.46	1.36	1.37	1.27*	1.37	1.38

1.42　1.34　1.43　1.42　1.41　1.41　1.44　1.48　1.55
1.37

为了方便通常把频数分布表和频率分布表用同一张表给出．列表的具体步骤如下：

(1) 找出数据中的最大值和最小值．此例中最大值为1.55，最小值为1.27.

(2) 确定数据的极差．所谓极差就是数据中最大值与最小值之差，用 R 表示．此例中 $R=1.55-1.27=0.28$.

(3) 决定组距和组数．在样本数据较多时通常分成10～20组，在样本数据少于50时分成5～6组．先决定组距，然后决定组数．组距就是每组中最大值与最小值之差．组距决定于极差 R．此例中 $R=0.28$．看来组距以0.03较好．这样可等距地分为10组。

(4) 决定分点．如果我们按原来纤度取值的精度分组，就要分为1.26～1.29，1.29～1.32，…，1.53～1.56共十组．但这样分组时，对纤度恰好是1.29的数据是分到1.26～1.29这一组，还是分到1.29～1.32这一组呢？为了避免这个麻烦，只要使分点比原来数据精度高一位就可以了．就是下面的10组：

1.265～1.295；1.295～1.325；……；1.535～1.565.

(5) 数出频数．所谓频数就是样本中个体（数据）落在每个组内的个数．数频数可用选举时唱票的方法在表上直接算出．

(6) 计算频率．所谓频率就是频数与样本总数之比．通常用百分比（%）表示．

上述100个数据的频数与频率分布表如下：

表 1

分　　组	频　率　累　计	频　数	频　率
1.265～1.295	一	1	0.01
1.295～1.325	丅	4	0.04
1.325～1.355	正 丅	7	0.07
1.355～1.385	正 正 正 正 丅	22	0.22
1.385～1.415	正 正 正 正 丅	24	0.24
1.415～1.445	正 正 正 正 丅	24	0.24
1.445～1.475	正 正	10	0.10
1.475～1.505	正 一	6	0.06
1.505～1.535	一	1	0.01
1.535～1.565	一	1	0.01
Σ		100	1.00

(7) 作频数（或频率）统计图.

首先选取坐标系：在横坐标上标出每组的上、下限，在纵坐标上标出每组的频数/组距(或频率/组距)．然后以每组的组距为底边，以每组的频数/组距（或频率/组距）为高作矩形，就得到频数（或频率）分布图，也就是统计图．对维尼纶纤度的统计表可得如下的分布图：

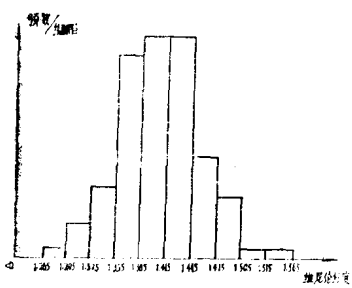

图 1　频数分布图

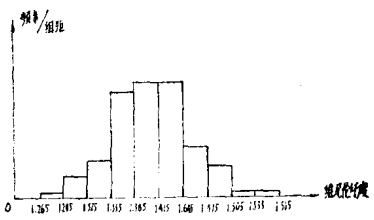

图 2　频率分布图

需要说明的是，上述分布图中的纵坐标为什么要取为频数/组距(或频率/组距)？这是为了使每组的矩形面积正好是每组的频数（或频率）而各矩形面积之和正好是总频数 n(或总频率 1)．频数分布图与频率分布图是完全相似的图形，后者比前者缩小 n 倍（n 为样本容量）．

（四）样本均值和样本方差

通过对实验数据列表和制图，使我们对实验数据的分布状况从整体上有了比较直观的认识．但是这还是不够的，因为隐蔽在数据背后的统计规律还不清楚．必须利用一定的数理统计方法才能表现和揭示出来．通常的方法是求出“表示”总体统计规律的特征数．为什么在“表示”二字上加引号呢？那是因为我们不可能直接求出总体的相应的特征数，只能通过样本的特征数来近似地表示总体的相应性质．

常用的特征数有两类．一类是描述数据集中性质的特征数，这种特征数常用的有平均数，中位数，众数等．另一类是描述数据离散性质的特征数，这种特征数常用的有方差、标准差、平均差、极差等．这里只介绍平均数(算数平均数)和标准差两个特征数．

1. 样本均值

样本均值是表现样本中数据集中性质的各种平均数中最基本的一种；它就是算术平均数．均值能反映出样本的平均水平，用 $\bar{x}$ 表示．

对于不分组数据来说，

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,$$

其中 n 为样本容量，$x_i(i=1,2,\cdots,n)$ 为实验数据。

对于分组数据来说，设有 N 个数据已分为 n 组，并以 x_1，x_2，…，x_n 表示各组组中值①，f_1^*，f_2^*，…，f_n^* 表示各组的频数，将属于同一组的数据都用该组的组中值代替，这时有

$$\begin{aligned}\bar{x}&=\frac{1}{N}(f_1^*x_1+f_2^*x_2+\cdots+f_n^*x_n)\\&=\frac{f_1^*}{N}x_1+\frac{f_2^*}{N}x_2+\cdots+\frac{f_n^*}{N}x_n.\\&=f_1x_1+f_2x_2+\cdots+f_nx_n\\&=\sum_{i=1}^{n}f_ix_i.\end{aligned}$$

其中 f_i 为第 i 组的频率。

实际工作中只有在数据 x 不大又不小时，我们才按照上述公式来求均值，如果数据 x 很大（或很小）时应用公式求均值是很麻烦的。所以通常都对数据先实行变换，使数据简化。然后用简化后的数据来计算均值。下面介绍均值的两种简算法：

(1) 将每一实验数据 x 都减去同一常数 A，用 u 表示它们的差。即有关系式：

$$u_i=x_i-A\ (i=1,\ 2,\ 3,\cdots,n).$$

---

① 各组组中值就是各组的 $\frac{\text{上限}-\text{下限}}{2}+\text{下限}$. 也就是各组的中点.