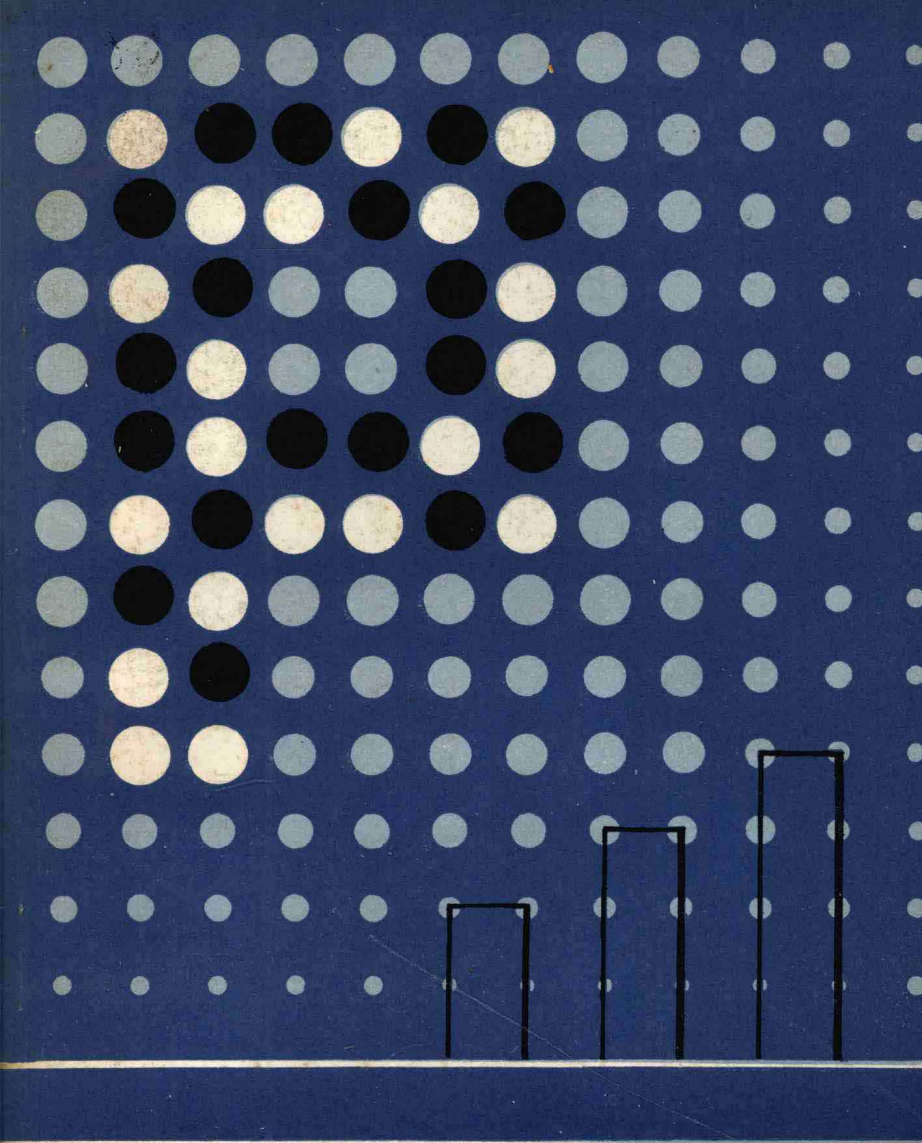

中学概率统计基础

江苏教育出版社

中学概率统计基础

苏起凡　金炳陶

江苏教育出版社

内 容 提 要

本书根据《中学数学教学大纲》编写，选材密切结合中学教材。前三章：样本分布——图表法，总和的记法，样本分析——特征数法，与初中教材“统计”一章平行；后三章：随机事件及其概率，概率的基本运算法则，随机变量（离散型），与高中教材“概率”一章平行。全书通俗易懂，并注重解题技能技巧的训练和指导，适合不同学制的中学数学教师教学概率统计时参考，并可作为中学生学习辅导。

本书前三章由苏起凡同志编写，后三章由金炳陶同志编写。

中学概率统计基础

苏起凡　金炳陶

江苏教育出版社出版

江苏省新华书店发行　海门印刷厂印刷

开本787×1092毫米　1/32　印张8.75　字数185,000

1984年10月第1版　1984年10月第1次印刷

印数 1—10,000 册

书号：7351·010　　定价：0.88元

责任编辑　何震邦

编 者 的 话

概率论与数理统计是研究随机现象统计规律性的数学学科，也是现代质量管理的重要工具。随着四化建设的发展，应用日益频繁。今天，在高等学校的理、工、医、农、经济等专业的教学计划中，它是一门必修课；在中学数学教材中，也有相应的内容。在中学里学习这方面的初步知识，对于毕业后参加工作或进一步学习都有重要意义。

初次接触概率统计的读者会有难教难学的感觉，这是毫不奇怪的。因为，这门学科的研究对象、思考方法与隶属确定性现象的算术、代数、几何、微积分等有较大的差异。学好概率统计，必须正确地建立新的概念，熟练地掌握各种运算法则。这需要有一个逐步培养和训练的过程。编写本书的目的在于满足中学数学教师和中学生教与学的需要，帮助他们解决难教难学的实际困难。编写时，力求紧扣中学数学教学大纲，结合中学数学教材，按照先直观后抽象，先简单后复杂的原则去组织材料。书中选取的例题和习题注重应用，重点讲清解题思路，致力于提高分析问题和解决问题的能力。

本书前三章，从实(测)数据的整理、分析入手，引入最基本的数据处理的统计方法，也为全书提供了必要的直观背景。后三章介绍概率论的若干基本概念以及在应用中占有重要地位的二项分布。

本书主要供中学数学教师教学参考和中学生课外阅读。也可供广大青年自学时参考。

由于水平有限，书中不妥和谬误之处恐所难免，恳请广大读者批评指正。

编 者

一九八四年元月

目　录

第一章　样本分布——图表法……1
§1　引　言……1
§2　总体、个体、样本(子样)……2
§3　顺序排列……4
§4　频数、频率分布表……7
§5　直方图……15
§6　累积频率直方图及累积频率分布……22
第二章　总和的记法……34
第三章　样本分析——特征数法……45
§1　引　言……45
§2　位置度量——平均数……45
§3　离散度量——样本方差、标准差……59
第四章　随机事件及其概率……78
§1　引　言……78
§2　事件的概念……82
§3　事件间的关系和运算……93
§4　事件的概率……110
第五章　概率的基本运算法则……143
§1　概率的加法公式……143
§2　条件概率与概率的乘法公式……161
§3　事件的独立性……172
§4　独立重复试验与二项公式……189
第六章　随机变量(离散型)……207

§1 随机变量的概念…… 207
§2 离散型随机变量的概率分布…… 212
§3 随机变量的数字特征…… 229
§4 二项分布…… 238
附录一 阶乘对数表…… 251
附录二 淘松分布数值表…… 254
附录三 习题、复习题解答与提示…… 257

第一章　样本分布 ——图表法

§1　引　言

在实际生活中，经常要做调查工作，收集很多数据。例如，人口普查、产品质量情况调查等等。人们常把数据的登记、画表等工作称为**统计**。这种统计有时能粗略地定性地反映出一些问题，但往往是比较粗浅的。随着社会生产和科学技术的不断发展，对事物性质的判断已不能满足于粗略的估计和定性的分析，而希望从表面上看起来是杂乱无章的大量数据中，运用科学的方法，从定量的角度对事物的性质作出论断。统计学就是解决上述问题的一个很有成效的工具，它是数学的一个分支，且有它自身的方法和理论的。这里所说的统计学就是数理统计学的简称。

收集和积累数据是统计的基础和依据。必须深入实际、深入现场，做好观察和科学试验工作，以求得第一手的真实、准确、可靠的数据资料。切不能粗枝大叶、马虎敷衍，更不能事先作出结论，有目的地去挑选有关的数据或者篡改数据，甚至用虚假的方法来证明先前的结论。这样做，会给我们的事业带来不良的后果。

本书统计部分的内容主要是对众多的数据进行整理，绘制成各种图表，或者由观测数据来计算各种统计量。这些图表

和量能够初步定量地反映事物内部的规律及它的主要特征，但这样的统计还是属于描述性质的，因此人们称它为**描述统计**。要想对事物的本质和规律性有更深入的了解，必须对数据进行统计分析，并对事物的总体作出推断，这方面的内容在统计上称为**统计推断**。由于它超越了中学数学的范围，本书不再讨论。

§2 总体、个体、样本（子样）

统计学，把研究对象的全体称为**总体**。例如，电视机厂要考察某月生产的电视机的质量情况，那么该月生产的全部电视机就构成一个总体；又如，在研究南京气温变化情况时，南京每天的平均气温的全体也构成一个总体。总体中的每一个元素（或基本单位）称为**个体**。如，上面两个例子每一台电视机和每一天的平均温度都分别是上述各个总体中的个体。

如果总体只包含有限个个体，那么称该总体为**有限总体**；包含无限个个体的总体称为**无限总体**。例如，某月生产的电视机总台数是一个有限数，该总体就是有限总体。又如，研究某试验田的棉花纤维长度时，全部棉纤维的根数是难以一一数出来的，这样的总体便是无限总体。再如，南京的每天的平均气温，如果从有记录的那年开始直至如今，其天数是很可观的，我们把这样的总体也近似地当作无限总体。有些工业自动化程度很高，产品又源源不断地生产出来，这种产品的全体也可以理解为无限总体。

如果从总体中抽取一部分，那么这一部分就称为总体的一个**样本**（或叫子样）。例如，从某月生产的电视机中任意抽

取20台进行质量检验，那么这20台电视机就是一个样本。这个样本包含了20个个体，我们又称这20台电视机是**容量**为20的样本。这里容量的意义就是一个样本里所含有个体的个数。通常把容量达到或超过50的样本称为**大样本**，而把小于50的称为**小样本**。例如，糖厂用自动打包机将糖装入蒲包中，每包标准重量是100斤，每天开工后需要检验一次打包机工作是否正常，某日开工后从200包中任抽 9 包其重量如下：（单位：斤）

99.3	98.7	100.5	101.2	98.3
99.7	99.5	102.1	100.5	

这 9 包称为容量为 9 的样本。

现代化大生产，产品的批量往往很大，如螺丝钉、螺栓等，每天可以生产成千上万个，如果逐个检查将会花费很多的人力、物力和时间。又由于对某些产品的检验是带有破坏性的，如电灯泡、显象管的寿命，砖的强力，布的拉力等等，如果采用全部检验，势必造成大量毁坏。

基于上述两种情况，我们总是希望只选取少量的产品进行研究，即由研究样本的性质，来估计或者判断总体的性质。这反映了局部与全体之间的辩证关系。

要使样本的性质能够充分地反映总体的性质，对样本的选取就有一定的要求：一是样本中各个个体的选取必须具有代表性。因此研究者不能先抱成见，随意把合乎自己意图的个体留下，不合乎意图的删去。二是样本中各个个体的选取必须是独立的。也就是说，各次选取的结果互不影响。

以上两条在实践上是经常可以做到的。不妨以生产螺丝钉为例，只要我们生产的条件（车床、原材料、操作人员、规格

标准……)稳定不变,且样本又是任意选取的,那么,这种样本就具有代表性。又由于产品很多,可近似地当作无限总体,后选到的螺丝钉并不受先取到的影响,从而独立性也可以得到保证。

现以糖厂自动打包机为例,用下图表示总体、样本、数据之间的关系。

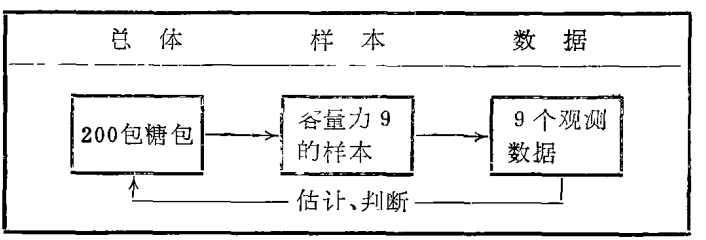

图1-1

§3 顺序排列

统计学的基本任务就是**由样本的规律性来估计、判断总体的规律性**。样本的规律性和大量的观测数据所呈现出的各种现象是有着密切关系的。现在通过下面的实际问题来说明。为了了解某品种棉花的质量情况,从棉包中任意选取106根棉纤维,其长度(单位:毫米)列表如下:

表1-1 **106根棉纤维长度表**

29.99	29.74	28.46	31.50	30.57	28.89
29.26	30.89	30.27	29.47	30.62	27.82
31.71	28.61	28.96	30.97	30.62	29.58

29.33	30.12	30.41	30.24	31.15	28.04
28.68	30.74	30.74	29.06	29.38	31.16
29.91	29.66	28.22	31.32	28.73	29.21
30.75	30.22	29.67	29.97	30.48	29.41
31.44	28.44	28.82	30.85	30.52	30.27
29.24	29.45	29.99	29.72	32.38	31.16
30.75	28.22	28.69	29.15	30.47	30.19
29.39	29.67	29.94	29.97	29.69	31.36
30.78	28.37	28.79	30.52	30.24	29.24
29.43	30.45	27.92	28.64	29.06	31.94
31.03	30.66	29.33	29.59	30.14	29.89
28.91	27.25	28.55	30.92	30.00	29.48
29.86	30.33	30.58	29.29	31.58	27.64
28.57	31.69	30.96	28.94	29.32	30.61
30.38	29.86	29.53	30.08		

我们最初接触这些资料时，很难看出这一样本具有什么规律性。如果反复、细心地观察，便可以发觉最短的长度是27.25毫米，最长的是32.38毫米，小于28毫米或大于32毫米的根数很少。但是还不能很快地说明落入长度相同的两个范围内的数据个数是否一样多，例如，现考察落入29.245至29.745范围内的数据个数和落入31.245至31.745范围内的数据个数是否同样多。为此，要对数据进行整理，使能逐步地看出其中的规律性。现在把表1-1按照从小到大的顺序排列起来，得表1-2。

表1-2　按顺序排列的棉纤维长度表

27.25	27.64	27.82	27.92	28.04	28.22
28.22	28.37	28.44	28.46	28.55	28.57

28.61	28.64	28.68	28.69	28.73	28.79
28.82	28.89	28.91	28.94	28.96	29.06
29.06	29.15	29.21	29.24	29.24	29.26
29.29	29.32	29.33	29.33	29.38	29.39
29.41	29.43	29.45	29.47	29.48	29.53
29.58	29.59	29.66	29.67	29.67	29.69
29.72	29.74	29.86	29.86	29.88	29.89
29.91	29.94	29.97	29.97	29.99	29.99
30.00	30.08	30.12	30.14	30.16	30.19
30.22	30.25	30.27	30.27	30.33	30.38
30.41	30.45	30.47	30.47	30.48	30.52
30.52	30.57	30.58	30.61	30.62	30.66
30.74	30.75	30.75	30.78	30.85	30.89
30.92	30.96	30.97	31.03	31.15	31.16
31.32	31.36	31.44	31.50	31.58	31.69
31.71	31.94	32.24	32.38		

由表1-2可知落入29.245至29.745范围内的数据要比落入31.245至31.745范围内的数据多。这里需要说明一点，当样本的一批观测数据其数值的大小与出现的先后次序有关时，就不应该不顾出现的先后次序而将它们按大小顺序重新排列。例如，研究南京的年平均温度的变化，那么，各年都有一个温度与其对应，即年平均温度同所在的那一年有关，此时数据就不能按大小顺序排列，否则就看不出不同的年份的年平均温度变化的情况。又例如，某台车床加工一种零件时，因为车刀不断地磨损，使这台车床加工的零件尺寸受到车刀的磨损影响，因而零件的尺寸与加工的先后次序有关，如果不顾加工的先后次序，而将零件的尺寸按从小到大的顺序排列起

来，那么车刀的磨损对于零件尺寸的影响就显示不出来了。

§4 频数、频率分布表

从§3可知，由原始的数据表1-1转换成按顺序排列的表1-2，固然有其优点，但数据越多，工作量越大，甚至会出现差错，那么能否在原始的数据表的基础上来寻找样本的规律性呢？本节介绍的频数、频率分布表就是一个较好的方法。就以研究棉纤维长度的分布为例进行讨论。

由表1-1可知，最短的棉纤维长度是27.25毫米，最长的是32.38毫米，它们的差(即最大值－最小值：32.38－27.25＝5.13)称为**极差**或称**全距**。它给出了观测值的变化范围，在这个范围内数据的分布，一般来说是不均匀的，也就是说在具有同样长度的不同间隔内占有数据的个数并不相等。所以，我们常要对观测值的全距进行分组，以便深入地研究数据的分布情况。

如何分组和分多少组？这对研究分布将起着重要的作用。为了方便起见，实用时常采用**等距分组**，就是说将全距分为距离相等的若干组，这个距离又称为**组距**。每一组的左端点叫做**组下限**，右端点叫做**组上限**，两个端点统称**组限**。每一个组的中点称为**组中值**，它的计算公式是：

组中值＝$\frac{\text{组下限}+\text{组上限}}{2}$。而组距的计算公式是：组距＝组上限－组下限。或者是：组距＝后一组的组中值－前一组的组中值。

分多少组，目前没有统一的方法，只有一个基本的原则：

大样本通常分成10—20组比较好；小样本分成5—6组为好。在实际分组时可参考下表1-3。

表1-3 组 数

数据个数	组 数
50以下	5—6
50—100	6—10
100—250	7—12
250以上	10—30

极差、组数、组距三者之间的关系可用下式表示：

$$\text{组距} = \frac{\text{极差}}{\text{组数}}。$$

为了计算上的方便，组距常取测量单位的整数倍。

表1-1给出了106个数据，参照表1-3确定分为11组，又已知极差是5.13毫米，从而可确定组距的值是$\frac{5.13}{11}\approx 0.5$。

下面就要计算各组分点(组限)的数值，这里分点的选取必须恰当，否则会产生一些不合理的现象。比如，以最短长度27.25毫米作为第一组的组下限(第一分点)，那么组上限应为27.25＋0.5＝27.75毫米，这样第一组就是27.25—27.75，第二组是27.75—28.25；第三至十一组分别是28.25—28.75，28.75—29.25，29.25—29.75，29.75—30.25，30.25—30.75，30.75—31.25，31.25—31.75，31.75—32.25，32.25—32.75；再观察106个数据，可以发现其中的30.25和30.75都是组限，那么30.25该属于29.75—30.25、30.25—30.75这两个组中的哪一个组呢？同样的道理，30.75也不好判别该属于第七组还是第八组。为了避免这种麻烦，就应重新规定分点选取的方

法，以明确地表明每一个数据该属于那一组。平时常采用下面的两种方法：

一是分点比观测值的精度高一位。通常规定如下：当观测值精确到$\frac{1}{10}=0.1$时，那么分点就可以精确到$\frac{1}{20}=0.05$；同理，如果观测值精确到$\frac{1}{100}=0.01$，那么分点可精确到$\frac{1}{200}=0.005$。依据这个原理将106个数据仍分成11组，组距仍为0.5，由于观测值精度是0.01，分点可精确到0.005，为了使最小值27.25落在第一组内，所以第一组的第一个分点（组下限）应是27.25－0.005＝27.245，第一组的第二个分点（组上限）应是：组下限＋组距＝27.245＋0.5＝27.745，第二组的组下限与第一组的组上限相同，第二组的组上限＝第二组的组下限＋组距＝27.745＋0.5＝28.245，按照上述方法，不难得到其余各组的分点，分得的十一组情况如下：

27.245—27.745　27.745—28.245　28.245—28.745
28.745—29.245　29.245—29.745　29.745—30.245
30.245—30.745　30.745—31.245　31.245—31.745
31.745—32.245　32.245—32.745

依据这种确定分点的方法，30.25应属于第七组，而30.75应属于第八组。

二是以最小值27.25作为第一组的组中值，这样第一组的组下限就是最小值减去组距的一半，即有$27.25-\frac{0.5}{2}=27.0$，组上限＝组下限＋组距＝27.0＋0.5＝27.5，但组上限规定它不属于第一组，所以用27.5以下表示真正的第一组的组上限。这样第一组是27.0—27.5以下，第二组是27.5—28.0以下，其

余类推。有时为方便起见这十一组可以简记如下：

1. 27.0—　　2. 27.5—　　3. 28.0—　　4. 28.5—

5. 29.0—　　6. 29.5—　　7. 30.0—　　8. 30.5—

9. 31.0—　　10. 31.5—　　11. 32.0—

这两种常用的确定分点的方法，本书称前者为第一法，称后者为第二法。这两种方法都是很重要的。它们之间互有关系，可以互相转换。现举例说明。

例1　下面三组是采用第二种分点法得到的，第一组42.5—　，第二组44.5—　，第三组是46.5—　，试改用第一种分点方法，将所得的三组写出来，并计算这些组的组中值。

解　由题设可知组距等于2.0，由第二法可知它的第一组的组下限等于数据的最小值减去组距的一半，所以，数据中的最小值等于组下限加上组距的一半。即有最小值＝组下限$+\frac{\text{组距}}{2}=42.5+\frac{2.0}{2}=43.5$，再由数据精确到$\frac{1}{10}=0.1$，那么，由第一法确定的分点可精确到$\frac{1}{20}=0.05$，为使最小值43.5落在第一组内，所以由第一法得到的第一组的组下限是43.5—0.05＝43.45，组上限＝43.45＋2.0＝45.45。这样改用后的三组分别是43.45—45.45　　45.45—47.45　　47.45—49.45，组中值分别是$\frac{43.45+45.45}{2}=44.45$，

$\frac{45.45+47.45}{2}=46.45$，　$\frac{47.45+49.45}{2}=48.45$。

当分组的组数、组距以及各组的分点确定以后，就要分别确定落入各组的数据个数，在统计上称这个数为**频数**(次数)。频数与样本容量的比值称为**频率**。如果数据已经按顺序排

列，那么各组的频数、频率就可以很快地计算出来。例如，第一组的频数是2，它的频率就是$\frac{2}{106}\doteq 0.019$。表1-4就是106根棉纤维长度的频数、频率分布表。

表1-4　频数、频率分布表

组　号	分　　组	频　数	频　率
1	27.245—27.745	2	0.019
2	27.745—28.245	5	0.047
3	28.245—28.745	10	0.094
4	28.745—29.245	12	0.113
5	29.245—29.745	21	0.198
6	29.745—30.245	18	0.170
7	30.245—30.745	17	0.160
8	30.745—31.245	11	0.104
9	31.245—31.745	7	0.066
10	31.745—32.245	2	0.019
11	32.245—32.745	1	0.010
合计		106	1

如果数据不按顺序排列(今后讨论的就是这种情形)，那么计算频数的最快的方法，就是一一的读数据表，每读到一个数据，就在它所属的那个组旁边做一个记号，并累计这些记号，就得到各组的频数。形象地说，就是用选举时唱票的办法，数出落入各组的频数。下面的表就是未按顺序排列的数据的频数、频率分布表。表1-5中最后两列是为以后用的，它的概念和计算将在本章§6介绍。