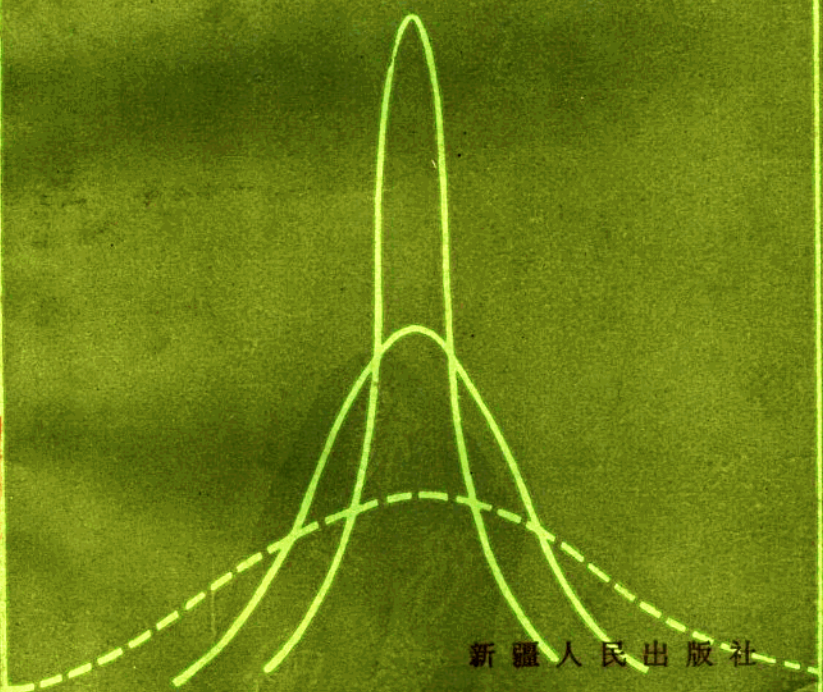
罗崇谢 编著

概率论与数理统计自学教程

下 册



新疆人民出版社

目　录

第二部分　数理统计

在上册中讲述了概率论的基本内容，那些内容是数理统计不可缺少的理论基础．概率论是数理统计的理论根据，而数理统计则是概率论的实际应用，“概率论”与“数理统计”是彼此联系互为依存的姐妹科学，因此往往将“概率论”和“数理统计”看成是，“概率论与数理统计”这门完整科学的两个部分，即理论部分和应用部分．但应指出在理论部分中也有应用，在应用部分中也有理论，不可截然分开．

在概率论中我们采用了分布函数来描述随机变量的变化规律，但其某个随机变量的分布函数的确定，都必须对随机变量进行大量试验和观察，取得大量数据，然后对这堆数据进行审定、整理、分析，从而对其某些性质进行估计、假设和分析推断．数理统计是运用概率论的基本知识，对要研究的随机现象重复进行大量次的试验，研究如何合理地获得数据资料，建立科学而有效的数学方法，根据对所获资料整理和分析的结果，对我们感兴趣的问题（如概率分布、数学期望、方差等概率特征）作出充分合理的估计和推断．这就是数理统计学的重要分支——统计推断要解决的问题，除此而外，数理统计中还有多元分析、试验设计、抽样理论等重要分支．

第七章　数理统计的基本知识和抽样分布定理

§7.1　数理统计

一．数理统计的内容和方法

1．数理统计的基本内容

①表达一件事物的性质

如某种动物的平均寿命，某批棉花纤维的平均长度，某射击试验中误差大小的规律分布等，这些问题都属估计问题．其中包括参数估计和非参数估计，或从性质分有参数的点估计，参数的区间估计，分布的估计以及适应性估计等不同内容．其中大部分内容将在第八章**参数估计**中进行讨论．

②比较两件事物的差异

如某生物种是否来自已知种群，某批机器零件是否来自某车间，工艺改革前后的产品是否有明显的差异等，都是属于这类问题．它在数理统计中被称为**假设检验**，并分为参数性的假设检验和非参数性的假设检验．参数的假设检验主要是对正态母体中数学期望和方差的检验，非参数检验包括，分布函数的假设检验，独立性假设检验，适合度检验等，其详细内容将在第九章**假设检验**中给予讨论．

③分析影响事物变化的因素

如某生物种群的生境条件是由多个因素构成的，每个因素的改变，都或多或少的对生物种群的生长和繁殖带来一定的影响，但我们要问，究竟是哪个（或哪几个）因素的影响

最大，是主要因素，而哪个（或哪几个）因素影响较小，是次要因素。也就是说，要区别各因素对某种群生长繁殖的影响是系统性的还是随机性的。再如在同一台车床上生产出来的两批另件其精度是有差异的，但这种差异可能是由于原料的不同，或加工处理方法不同，或由技术水平不同的工人生产，或由原料的不均匀性和周围环境的微小改变所致。究竟是哪种情况起主导作用呢？这些问题将在第十章**方差分析**中予以讨论。

④一种事物的两种性质之间，或环境对事物性质的关系

如某一年龄的男子的身高与体重之间的关系，某植物的生长与温度湿度之间的关系，青砖的抗压强度与抗折强度之间的关系，某纺纱车间的湿度与细纱断头率之间的关系等等。上述诸例的两种性质或事物性质与环境之间，在客观上都是有联系的，但它们之间的联系不是必然性的联系，它们之间的关系不是普通的函数关系，而是相关关系。如身体高的男子一般都比身体低的男子的体重重一些，但也有例外。这类关系问题在数理统计学中称为“相关”问题，相关关系是不同与函数关系的，这些问题将在第十二章**回归分析**中予以讨论。

⑤研究取样与试验方法

统计推断的基本思想就是从一个庞大的总体（有限或无限）中，抽取一个容量大小为 n 的样本，通过对样本的种种性质的分析与研究，用其研究结果去估计、推断和臆测总体的相应性质。因此就产生了选取多大的样本？如何选取样本？这类问题属于抽样方法问题，将在第十一章**抽样试验**中予以讨论。

在生产和科研中，人们经常要进行实验或试验，那么怎样安排试验才是科学的呢？怎样才能在最少次的试验中得到

最大的信息呢？也就是说，如何科学地设计试验，又如何科学地分析试验的种种结果，从而为生产和科研提供可行的实施方案，并指出指标变化的趋向等，这类问题将在第十三章**正交试验设计**中予以讨论。

上面提及的五方面内容，是数理统计学处理和解决的基本问题，当然不是全部问题。

2．数理统计的基本方法与步骤

数理统计方法是数学方法与统计方法的兼用与结合，数学方法有：解析法、集合法、空间几何法，特征函数法及其它一些方法；而统计方法最基本的有：统计分组法、大量观察法和综合指标法等。

数理统计学的方法是用于不确定因素（这儿指的是随机因素）占优势的各个领域和问题中，进行统计推断和试验分析的重要工具。在用数理统计方法处理具体问题时，基本步骤是：*a*）收集，整理；*b*）分析；*c*）推断等三大部分。这三部分工作分别属于叙述性，分析性和推断性的工作。

数理统计学是一门应用很广的科学，它的思想、理论和方法，在促进工农业生产和科学技术的迅速发展中，起着巨大的作用。它是处理和解决随机现象数量规律的强有力的工具。在生物、地质、水文、气象、战争、通迅以及与人有关的公共事业中都存在着广泛的应用。它的思想、理论和方法与其它一些学科结合，又形成了多种专业性的数理统计，如：医学统计、纺织统计、工业试验统计、商业统计．水文气象统计……等。

二．数据整理

1．数据资料整理的意义

数据资料的收集与整理是数理统计中第一步工作，是获得信息的重要步骤，是分析推断的依据，同时也是分析和推断成功与否的关键，所以数据整理是数理统计中不可忽视的重要环节。

通过试验或观察所得到的数据，往往是零星的、孤立的、甚至是杂乱无章的。要想从这堆来之不易的数据中，提取我们所感兴趣的可贵信息，从而揭示出它们的内在联系及其变化规律性，就必须对原始资料进行科学而细致地整理。所谓整理就是从大量错综复杂的原始数据中，判定其真实可靠性，整序归类，使其系统化，以便于统计分析，使其能得出正确的，科学的结论。这种为了揭示随机现象的统计规律而进行的对数据加工整理手续，就称为**数据整理**。

2．*数据资料的分组整理法*

数据资料分组整理法，就是要按照一定的指标，将观察所得的数据分门别类地分成若干组，把同一现象，同一类型的数据进行合并，使它们与其它现象和其它类型区分开来。数据资料的整理分类是统计归纳的基础，是使资料系统化、规律化。从而能正确地反映出事物的本质和规律所必需。在数据整理时，一定要注意数据的完整性、真实性和准确性。因此，必须首先对原始数据进行严格地检查与核对，尤其对那些特大和特小的可疑数据，更应该认真地反复核对，力求数据确实可靠。

对原始数据检查核对时应注意：

(i) 数据本身是否有差错，要考虑记录是否完整无失、有无损坏和遗漏，有无笔误、虚构和人为的主观改动；

(ii) 取样是否合理。这里主要考虑有无取样过少和缺乏独立性和代表性（比如不全、不同质、不随机、不独立或具

有主观意向等）；

(iii) 数据合并是否合理。在数据整理时，对不同质不同类的数据不能随便合并。如对公畜与母畜，健康与有病的动物不能合并归类，另外对品种不同，年龄不同，胎次不同，营养不同、时间不同、试验因素不同等数据不应随便合并，如发现错并者应给予改正。

(1) 离散型变量数据资料的分组整理

对离散型随机变量而言，数据资料的分组整理是比较简单的，一般称为**单项式分组法**。现举例说明如下：

【例7.1】 某纺织厂在1972年10月纺制18号(32支) 棉纱过程中，连续进行100次生条手拣棉结的独立试验，其结果按试验的顺序列成表7.1，试对这100个数据进行整理。

表7.1 1972年10月生条棉结数（单位：粒/克）

9	8	4	7	6	7	10	7	8	3	8	10	8	7	7	6	8	9	8	6
6	8	7	6	6	8	7	5	11	10	8	5	5	9	4	8	6	5	10	7
12	8	11	5	6	6	8	8	3	7	6	9	5	12	9	4	8	8	5	4
6	9	6	7	9	11	7	12	6	9	12	11	4	8	10	8	4	10	8	4
4	7	10	3	7	6	3	7	5	4	5	7	4	8	8	8	8	6	6	3

在表7.1中每一个数据，是在所进行的100次独立试验中，每次试验的观察结果，称为**随机变量的观察值**。如表7.1中第一行的数据9，8，4，7，6，7，…分别是第1，2，3，4，5，6，…次试验的观察值。不同试验的观察值，在数值上可能是一样的，也就是说，不同次的试验可能有相同的结果。如表7.1中第4次试验和第6次试验的结果（观察值）都是7。而表7.1中所出现的不同的数值，称为**这批数据的变值**。在这里共有十个变值，可记为：

$$x_1=3,\ x_2=4,\ x_3=5,\ x_4=6,\ x_5=7$$
$$x_6=8,\ x_7=9,\ x_8=10,\ x_9=11,\ x_{10}=12.$$

即在此100次试验中，所出现的最小变值是$x_1=3$，最大变值是$x_{10}=12$。

在表7.1的100个数据中，变值$x_1=3$其出现了5次，占总次数100的5%=0.05，这里数字5称为变值3出现的频数，而0.05则称为变值3出现的频率，对其它变值也可同样理解。总之，在某一批观察数据中，变值x_i出现的次数称为变值x_i的频数，记为f_i^*，各变值的频数之和（总频数），即为这批观察数据的总个数n（此处$n=100$），变值x_i的频数f_i^*与数据总数n的比f_i^*/n，称为变值x_i的频率，并记为

$$f_i=f_i^*/n \tag{7.1}$$

不难看出：

(i) 任一变值x_i的频率f_i皆有：

$$0\leqslant f_i\leqslant 1,$$

(ii) $\sum_{i=1}^{m} f_i=1$。

其中m为该批数据中变值的个数。

在审查所得数据确为无误的条件下，可按数据整理的第一步工作，以变值为指标，将具有相同变值的观察值归为一类，并计算其相应的频数及频率，则得如下的频率分布表（如表7.2）。

由表7.2可见，生条棉结数ξ是一个随机变量，它的大致分布可由该表近似给出，即随机变量取值范围在3～12粒之间，且取值在6～8粒/克的可能性约为52%。

同样地，将该厂1973年1月份连续抽取100次的试验所

表7.2 生条棉结的频率分布表（1972年10月）

变值 X_i	3	4	5	6	7	8	9	10	11	12	总和
频数 f_i*	5	10	9	15	15	22	9	7	4	4	100
频率 f_i	0.05	0.10	0.09	0.15	0.15	0.22	0.09	0.07	0.04	0.04	1

得数据进行同样的整理，类似地有频率分布表（如表7.3）。

表7.3 1973年1月的生条棉结频率分布表

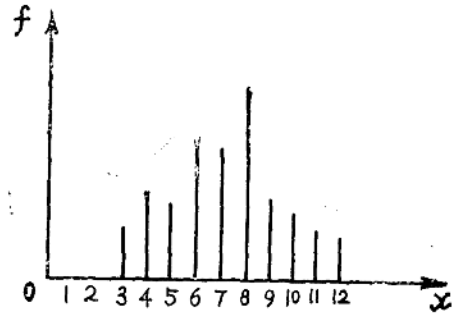
x_i	1	2	3	4	5	6	7	8	9
f_i*	7	12	18	17	20	13	6	3	4
f_i	0.07	0.12	0.18	0.17	0.20	0.13	0.06	0.03	0.04

由表7.3近似给出随机变量ξ(生条棉结数)，取值1～9粒之间，且取值在3～5粒/克的可能性约为55%。所以，比较表7.2和表7.3从总体上来说，1973年1月份的生条棉结数减少了，也就是说，生产质量有所提高。

若将分布表7.2和7.3表示为平面图形，则可选x轴为变值坐标，y轴为对应变值的频率坐标，其频率分布针形图是由一组各对应变值垂直线段组成（如图7.1）。

以上两例均为单项式分组法，其特点是以每个变值为一组进行合并统计。当然也不是说凡是离散型随机变量的数据资料都采用这种单项式分组法，若变量值很多时，仍用这种单项式分组法，则必然组数太多而每组内所含观察值太少，

因此不易发现其数据资料的规律性。对这种情况，多半将几个临近的变值并为一组，这样可减少以组数，显示规律，为进一步统计分析带来方便。



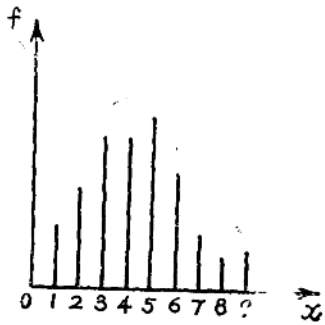


图7.1频率分布针形图

(2) 连续型变量的数据资料的分组整理

对连续型变量的数据资料的分组整理法，一般采用**组距式分组法**，其主要步骤是：

(i) 按照试验或观察的次序记录其观察值，即得原始数

据表；

(ii) 按上述三个方面审查核实原始数据；

(iii) 根据数据资料的多少确定分为几组，为使数据资料不发生重复和遗漏现象，一般采用前后相接的半开半闭区间。其分组原则是组数不宜太多也不宜太少，一般在实用上可参照下表进行。

表7.4 数据资料分组参照表

数据总数	可分组数
40～60	6～8
60～100	7～10
100～200	9～12
200～500	12～17

数据资料总数若比500还多时，也可适当增加组数，但一般以不超过20为宜。表7.4是一种习惯用法的概括，虽没有什么严格的理论根据，但在实用上是令人满意的。

(iv)找出最大值与最小值，算出极差，即

$$\text{极差}=\text{最大值}-\text{最小值}, \tag{7.2}$$

极差也称全距，它表示该组数据资料差异的最大幅度。

(v)由组数和极差确定组距。在实用中组距往往都采用相等组距，当然组距不一定是整数，但为了使用组距尽量简单些，可以适当地增大点数据资料的最大幅度。组距、组数和极差的关系为：

$$\text{组距}=\frac{\text{极差}}{\text{组数}}. \tag{7.3}$$

(vi)求出组中值，用组中值去代替组中所有观察值参加分析统计，此步骤可减少实际观察资料的变值数，为了分析上及数学处理上的方便，组中值应尽量选取简单，当然最好是整数。只要组下限和组距确定了，则

$$\text{组中值}=\text{组下距}+\frac{1}{2}\text{组距}. \tag{7.4}$$

由上面公式可知，只要确定第一组（即观察数据最小的一组）的组下限，那么第一组的组中值便可以求出，从而第二组的组中值即为第一组的组中值加上组距。一般地，第 i 组的组中值为第一组的组中值加上$(i-1)$倍的组距，这样所有的组中值都可算出。

(vii)确定第一下限及分组。分组时为了避免第一组中观察值个数过多，一般第一组的组中值最好选择接近或等于数据资料中的最小值。

(viii) 计算各组频数、频率、列出频率分布表。

(ix)由频率分布表给出频率分布图(折线图或直方图)。

【例7.2】 在20天内，从某维尼纶厂正常生产时生产报表上看到的维尼纶纤度（表示纤维粗细程度的一个量）的情况，有100个数据如表7.5，试进行整理。

解 因为维尼纶纤度是一个连续型的随机变量，虽然表7.5给出的是1.27～1.55中100个离散的观察值，其不同的变值仅有23个。但是当你观察另外100个数据，或测量精度有所提高（如精确到小数点三位）时，所得观察值中的变值可能不是上面的23个，而可能有增有减，甚至全不相同，所以我们不能用处理离散型变量数据的单项式分组法去整理，而应该用组距式分组法整理。

若对上述数据经审核无误后，参考表7.4将上面100个数据为分10组，因为该组数据中最大值是1.55，最小值为1.27，

表7.5 正常生产时维尼纶纤度原始数表

1.36	1.49	1.43	1.41	1.37	1.40	1.32	1.42	1.47	1.39
1.41	1.36	1.40	1.34	1.42	1.42	1.45	1.35	1.42	1.39
1.44	1.42	1.39	1.42	1.42	1.30	1.34	1.42	1.37	1.36
1.37	1.34	1.37	1.37	1.44	1.45	1.32	1.48	1.40	1.45
1.39	1.46	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

故极差为

$$1.55-1.27=0.28,$$

由组距计算公式得

$$\text{组距}=\frac{1.55-1.27}{10}=0.028.$$

为了计算上的方便和组中值尽量简单，我们取组距为0.03，且选1.265为第一组的下限，从而各组的组中值分别为

1.28, 1.31, 1.34, 1.37, 1.40, 1.43, 1.46, 1.49
1.52, 1.55

然后从所取观察数据出发，统计出现在各组中观察值的频数，同时计算其相应的频率，即可得频率分布表（如表7.6）；

由上面频率分布表，能够比较清楚地看出数据波动的规律。为了更进一步直观起见，可从频率分布表出发绘出其频

率分布图。频率分布图一般有折线图和直方图两种。

表7.6 频率分布表

分　组	组中值	频数统计	频数	频率
1.265～1.295	1.28	一	1	0.01
1.295～1.325	1.31	正	4	0.04
1.325～1.355	1.34	正正	7	0.07
1.355～1.385	1.37	正正正正正	22	0.22
1.385～1.415	1.40	正正正正正	24	0.24
1.415～1.445	1.43	正正正正正	24	0.24
1.445～1.475	1.46	正正	10	0.10
1.475～1.505	1.49	正一	6	0.06
1.505～1.535	1.52	一	1	0.01
1.535～1.565	1.55	一	1	0.01
Σ			$n=100$	1.00

(i) 折线图

折线图绘制方法是：以每组的组中值为横坐标 x，每组出现观察值的频率为纵坐标 y，这样每一组就对应着平面直角坐标系中一个点 (x, y)，将所得各点用折线连结起来，即为所求频率分布折线图，由表7.6可得图7.2。

(ii) 直方图

直方图的绘制方法是：以每组的组限为横坐标 x，每组出现观察值的频率为纵坐标 y。以每组的组距为底，以相应的

$$\text{频率密度}=\frac{\text{频率}}{\text{组距}}$$