

选煤实用数理统计

张荣曾 编

煤炭工业出版社

87
TD94
15
3

选煤实用数理统计

张荣曾 编

煤炭工业出版社

内 容 提 要

本书系统地介绍选煤生产、实验数据分析和设计中常用的数理统计方法和原理。

主要内容有：概率基础，随机变量的数字特征，样本的分布与数字特征，统计量分布与参数的区间估计，统计检验，样本数据异常的检验与质量管理图，方差分析，正交试验，相关与回归分析，取样与检验。

全书力求作到通俗易懂，每章都有应用实例；对各种数理统计方法所依据的数学原理，尽力运用比较简单的数学方法给出证明与阐述。

本书可供具有一定数学基础的选煤、选矿专业技术人员阅读，其他专业技术人员和大专院校师生亦可学习参考。

责任编辑：李 钟 奇

选煤实用数理统计

张 荣 曾 编

*

煤炭工业出版社 出版

(北京安定门外和平北路16号)

煤炭工业出版社印刷厂 印刷

新华书店北京发行所 发行

*

开本850×1168¹/₃₂

印张15¹/₄

字数405千字

印数1—1,600

1986年7月第1版

1986年7月第1次印刷

书号15035·2655

定价2.80元

前 言

数理统计方法是生产管理与科学试验中的有力工具。它的重要性与实用性愈来愈为人们所认识。过去，一般工科院校很少讲授这方面的内容，现在已成为绝大多数专业的必修课程。与此同时，社会上许多行业都相继开展了普及数理统计方法的活动。作者从一九七三年以来，从事数理统计方法在选煤中应用的普及与推广工作，深感需要一本结合选煤实际，系统地介绍常用数理统计方法的自学用书，供具有一般数学基础的现职工程技术人员使用。本书就是按照这个宗旨，在原教学讲义的基础上，参考已出版的同类书籍改编而成的。

本书除介绍常用数理统计方法及应用实例外，为了使读者进一步了解各种数理统计方法所依据的数学原理，尽力应用比较简单的数学方法进行证明与阐述。

本书编写过程中，得到煤炭工业部加工利用局、邯郸选煤厂和有关同志的帮助，在此一并致谢。书中难免有谬误之处，敬希读者批评指正。

作 者

1982年于北京

目 录

第一章 绪论	1
1.1 数理统计的研究对象	1
1.2 统计术语	4
1.3 常用数理统计方法	7
第二章 概率基础	11
2.1 随机事件的概率	11
2.2 事件间的基本关系	14
2.3 排列与组合	18
2.4 概率加法定理	22
2.5 概率乘法定理	26
2.6 全概率公式与逆概率公式	30
第三章 随机变量的分布	36
3.1 离散型随机变量的概率分布	36
3.2 连续型随机变量的概率密度函数	38
3.3 随机变量的分布函数	43
3.4 多维随机变量的分布	47
第四章 随机变量的数字特征	64
4.1 数学期望	64
4.2 方差	71
4.3 有关数学期望与方差的运算	79
4.4 矩、矩母函数	83
第五章 常用分布	92
5.1 二项分布	92
5.2 泊松分布	96
5.3 正态分布	101
5.4 中心极限定理	112
第六章 样本的分布与数字特征	118
6.1 切比雪夫不等式与大数定律	119

6.2	样本分布函数与直方图	123
6.3	样本均值与中值	127
6.4	有限母体样本均值的方差	130
6.5	样本方差与极差	133
6.6	样本均值与方差的简算法	139
第七章	统计量分布与参数的区间估计	146
7.1	统计量u分布	147
7.2	统计量 χ^2 分布	149
7.3	统计量t分布	157
7.4	统计量F分布	165
第八章	统计检验	170
8.1	引言	170
8.2	正态母体真值 μ 的检验	173
8.3	成对数据的比较	176
8.4	成组数据的比较	178
8.5	母体方差 σ^2 的检验	184
8.6	两母体方差的检验	187
8.7	符号检验法	191
8.8	秩和检验法	194
8.9	分布的假设检验	199
8.10	联立表独立性检验	206
第九章	样本数据异常的检验与质量管理图	211
9.1	可疑数据的判别	211
9.2	系统误差的检验	218
9.3	质量管理图	226
9.4	质量管理图稳定状态的判断规则	234
第十章	方差分析	245
10.1	引言	245
10.2	单因素方差分析	246
10.3	两因素方差分析	256
10.4	系统分组的方差分析	279
第十一章	正交试验	288
11.1	引言	288

11.2	拉丁方试验设计	289
11.3	正交表	300
11.4	正交试验结果的分析	308
11.5	有重复试验(取样)正交表分析	314
11.6	部分追加法	318
11.7	最优效果与缺失数据的估计	327
第十二章	相关与回归分析	332
12.1	引言	332
12.2	一元线性回归	334
12.3	回归方程的显著性检验	339
12.4	有重复的一元线性回归	348
12.5	回归方程的预报与控制精度	353
12.6	两条回归线的比较	357
12.7	曲线直线化的回归	361
12.8	多项式回归与多元回归	366
12.9	多元线性回归方程显著性检验	376
12.10	多元线性回归方程的预报精度	382
12.11	每个自变量在多元回归中的作用与回归系数 显著性检验	384
12.12	线性方程组求解、求逆消元法	386
12.13	逐步回归分析	393
12.14	逐步回归分析的计算程序	397
12.15	最小二乘曲线拟合的计算程序	405
第十三章	取样与检验	411
13.1	取样方案与概型	411
13.2	抽样检验	425
13.3	采样中系统误差的检验	430
13.4	质量检查各阶段误差的合成	433
13.5	确定单个子样总方差 σ_0^2 的试验方法	436
13.6	确定采样方差的试验方法	438
13.7	确定制样、分析方差的试验方法	441
13.8	质量检查结果准确度的检验方法	443
附 表		451

第一章 绪 论

1.1 数理统计的研究对象

在生产和科学实验中，我们经常要遇到很多数据，例如一批产品的数、质量指标，某种工艺过程或设备的效率、技术经济指标，等等。这些数据都是在一定的生产或试验条件下取得的。条件变了，数据也要随之而变。但是，在基本条件相同的情况下，它们也并不都是完全固定的。例如：按相同的采样、制样及化验规程操作，对一批煤炭的质量进行多次重复的测试，每次所得灰分值总是有高有低，参差不齐；再如一个选煤厂，在原煤、工艺设备和操作条件相同的情况下，所产精煤灰分，总的来说虽然是比较稳定的，但也难免会有波动。每次将取得什么结果确属难以预料，带有一定的偶然性。这类现象在日常工作中是不胜枚举的，它们的一个共同特点是：在基本条件不变的情况下，所进行的一系列试验或重复观察会得到不同的结果，每次将取得何种结果是不确定的，好像要随机会而定，这种现象称为“随机现象”。随机现象的每一种结果称为“随机事件”，它的取值称为“随机变量”。

为什么有些事情在基本条件不变的情况下会得出不同的结果？随机现象是否违背事物间普遍存在的因果关系规律呢？其实不然，世界上任何事物的发生都必然有它的原因，问题是这些原因往往是很复杂的，其中有些已为人们所认识，并为人们所掌握，在生产或试验中是可以控制的。通常所谓基本条件不变，也仅仅是指这些可以控制的主要条件未变。其实还有许多其它因素对试验结果也有影响，例如周围环境、气象条件，操作人员的情绪、心理状态间的微小差异和波动等对试验结果都会带来影响。

虽然它们不是影响明显的主要因素，每一种因素的影响也并不大，但是合起来也足以引起试验结果产生波动。由于这些因素中有些还没有被人们所认识和掌握，或者即使认识到也无法控制，有些因为它的影响太小，认为没有必要对它们加以控制，因此它们对结果的影响往往是随机的，波动的，有时大，有时小，表现出带有一种偶然性。通常将这些因素称为“随机因素”。

随机因素对结果的影响虽然表现出带有一种偶然性，但也并不是无规律可循。例如选煤厂在正常生产情况下产出的精煤灰分，确实参差不齐，是一个随机变量。就单次结果来看，很难发现有什么规律。但是如果我们统计了大量的数据，就可以发现其间的规律性仍然是十分明显的。表1.1.1中列出了三宝选煤厂连续

三宝选煤厂145批销售精煤灰分统计 表 1.1.1

灰 分 区 间 (%)	10.505 { 10.605	10.605 { 10.705	10.705 { 10.805	10.805 { 10.905	10.905 { 11.005	11.005 { 11.105	11.105 { 11.205	11.205 { 11.305
批 数	2	0	4	4	5	9	19	17
频 率, (%)	1.38	0.00	2.76	2.76	3.45	6.21	13.10	11.72
计 数	丁		正	正	正	正	正正正	正正正

灰分区间 (%)	11.305 { 11.405	11.405 { 11.505	11.505 { 11.605	11.605 { 11.705	11.705 { 11.805	11.805 { 11.905	11.905 { 12.005	12.005 { 12.105	12.105 { 12.205
批 数	19	20	16	10	7	5	2	2	4
频率(%)	13.10	13.79	11.03	6.90	4.83	3.45	1.38	1.38	2.76
计 数	正正正正	正正正正	一正正正	正正	丁正	正	丁	丁	正

145批销售精煤灰分的统计资料。从中可以看出,数据的分布很有规律,大量的数据集中在它的平均值11.37%附近,与平均值差值越大者出现的次数越少,差值越小者出现的次数越多;而且,数据均匀地分布在平均值的两侧,全部数据都落在10.50~12.20%区间之内。也就是说,虽然每次所产精煤灰分受随机因素的干扰很难断定一定会是什么结果,但是它却是有规律地在某一定的区间内围绕着一个平均值来回波动。很明显,如果生产管理得好,精煤灰份的波动就小,数据的分布就越集中;反之,波动就大,数据也越离散。这个例子说明,随机事件的发生也有它一定的规律。这种规律只有对随机事件进行大量重复的试验或观察后才能被发现,所以称为“统计规律”。

研究随机事件统计规律的学科就是数学中的“概率论”。数理统计则是利用概率论中有关随机事件分布规律的理论去整理、分析和研究生产试验数据的一门学科。它可以帮助我们从一些看来似乎是杂乱无章的一批数据中找出它的变化规律,去伪存真,由表及里地抓住事物的主要矛盾和发展趋势,并从中得出一些必要的有用结论。以上述三宝选煤厂销售精煤灰分为例,如果仅将各批灰分数据罗列出来确实是杂乱无章的,但经过整理和统计分析后,不但能看出它的分布趋势,而且可以预言,在正常生产情况下该厂销售精煤灰分不大可能(有95%把握)超出10.74~12.00%范围,如果发现某批超过此限,就意味着生产(或技术检查环节)中有某些异常,提醒人们去查找原因加强管理,这仅仅是一个极简单的事例。实际上在许多领域,特别是气象、水文、地质及地震预报等工作中,由于目前人们还不可能完全弄清楚其间确切的因果关系,还不可能用那些有确定性因果关系的物理定律对它加以概括、控制和预报,数理统计方法就显得特别重要。利用数理统计方法虽然不可能确切地指明一场灾害(如洪水,地震等)一定会在什么时候发生及其准确的规模有多大,但却可以推断它发生的可能性是多少,并能有一定的把握推断出它的规模大小。至于数理统计在选煤中的用途将在1.3节中作简要说明。

1.2 统 计 术 语

1. 母体、样本与个体

我们把在一次统计分析工作中所要观察和研究的对象的整体，称为母体(或总体)，其中每次考察的一个单元则称为个体。例如当我们要研究一个工厂的产品质量时，该厂所产每件产品的质量就是一个个体，全部产品的质量就是母体；当研究的是某一批产品的质量时，则该批产品的质量就是母体；当研究的是本月产品按批的平均质量时，则本月所产各批平均质量的全体就是母体，而其中每一批产品的平均质量则成为个体了。总之，什么是母体，什么是个体，并不是一成不变的。

要研究的是母体，但是母体通常很大，即使不大，有时也不可能对其中的每个个体都进行考察，因为有些检验是破坏性的，例如灯泡的寿命，煤炭的灰分等等。所以通常都是从母体中随机地抽取一部份个体进行考察，然后根据它们来推测母体的性质。这样从母体中抽出的一部份称为样本，样本中所包含的个体数目称为样本的容量(或大小)。数理统计的中心任务就是要解决如何根据样本去推断母体的问题。为了能够通过所得样本正确地去推断母体，样本必需有代表性，必须是从这个母体中随机抽取的，不能人为地从中挑选。数理统计以随机抽样为基础，如果抽样不是随机的话，任何数理统计方法和所得结论都将失效。

2. 随机误差与系统误差

数理统计的中心任务是要根据样本去推断母体。由样本来推断母体时，不可能完全没有误差，根据误差的来源或产生误差的原因，可将它分成两类：随机误差和系统误差。

由各种随机因素干扰所造成的样本观察值与母体真值间的误差，称为“随机误差”。由于周围事物中总难免存在着这样或那样一些尚未被认识 and 无法控制的随机因素，因此随机误差是不可避免的。随机误差的特点是有时大，有时小，有时正，有时负，它的大小和取向都不固定。在相同条件下，重复试验结果之间缺

乏重现性就是存在随机误差的表现。随机误差不可能完全消除，因为总会有些因素是人们所无法控制的，但是通过改善试验条件及精心操作可以使它减弱。运用数理统计方法处理试验数据，可以消除随机误差的干扰，并对它的影响作出定量的估计。

由那些作用明显、而且可以人为地调整控制的因素失调所造成的样本观察值与母体真值间的误差，称为“系统误差”。例如，试验用仪器、量具未校准，试验方法不正确，试验人员在操作中习惯性的偏向以及周围环境等外界因素的改变等，都可以引起试验结果不准确而带来系统误差。系统误差对试验结果的影响往往是恒定的，使试验结果恒定地比真值偏高或偏低。如果引起系统误差的因素本身也在变化，则系统误差当然也会随之而变，但这类变化通常很缓慢，且有一定的规律。通过重复试验和运用数理统计方法处理试验数据，也不能消除系统误差的影响，但是通过技术处理是可以消除的。

需要指出的是，系统误差与随机误差之间并不存在不可逾越的鸿沟，随着人们认识的加深和科学技术的发展，对于某些本来认为属于不可控制的随机因素，后来也可以得到控制并作为一种操作（或试验）条件规定下来。这时由这个因素失调而引起的误差就转变为一种系统误差。有时对于某些虽可控制但影响不大的因素在试验条件中并不纳入，这时它的波动所造成的误差就成为随机误差。

与随机误差和系统误差有关的两个常用术语是数据的“精确度”与“准确度”。精确度是指试验（或观察）结果间重现性的大小，重现性越好，数据的分布越集中，则精确度越高；反之，数据越离散，精确度也越低。精确度的高低只取决于随机误差的大小。准确度是指试验结果与母体真值的符合程度。它既取决于系统误差，也取决于随机误差。准确度好的，则精确度也一定高，但精确度高的不一定准确度就一定好。例如，在煤炭灰分化验工作中，如果精心操作，那么对同一煤样多次化验所得灰分值重现性就好，数据越集中，这时我们说化验工作的精度很高。但是如

果当时我们所用的天平没有校准，有系统误差，精度虽高也不能真实地反映所测煤炭灰分的大小，所以准确度并不好。

用数理统计方法从样本来估计母体时，一般只能根据样本的精确度来下结论。准确度如何则无法分辨。因此，对于精确度高的数据进行统计推断的结论，其精确度可能很高，这时可以根据样本对母体的真值作出一个区间很小的估计值。但这种估计未必就准确可靠，我们不能为这种假象所迷惑。统计推断是否准确，还要看观测中是否有系统误差而定。如果原始数据本身就不可靠，数理统计推断出的结论其准确性也必然不高。

3. 数据的类型

数理统计工作中所遇到的数据，按性质可区分为两大类，即计量值与计数值。计量值是指那些试验结果可在某个区间内连续取值的数据，例如测定一件物品的重量或长度，一台设备的处理量与效率，煤炭的灰分与水分等等。另外还有一类数据，如一批产品中次品的件数，产品上的缺陷数，每月事故的发生次数等等，这些数据有一个共同的特点，即它的取值不可能是连续的。一批产品中的次品件数只可能是1件，2件……等正整数，决不可能出现1.5件，1.26件次品。这类数据称为计数值。随机变量若是计量值，则称为“连续型随机变量”；若为计数值，则称为“离散型随机变量”。

对数据作这种划分在数理统计工作中是十分重要的，因为这两类随机变量的分布规律不同，数据处理时采用的统计方法也不相同。有些数据，如次品率，表面上看来很象计量值，因为它是一个百分数，可以在零至100%间取值，好像是连续的。实际并非如此。例如检查了N件，发现其中有n件为次品，则次品率为

$$p = \frac{n}{N}$$

由于次品的件数n是不连续的，故次品率p也不可能为连续。当N为100件时，p就不可能出现小于1%以下的位数。从这个例子可以总结出一个判别诸如百分率、比值等这类数据属性的方法，那就是要根据其分子的属性来区分。煤炭的灰分同样

也是用百分率来表示的，但是它的分子代表煤样灰化后的灰渣重量，是一个计量值，所以灰分也是计量值。计量数据有时也可以将它简化为计数值来处理。例如产品的质量指标往往是计量值，但是也可以按其质量区分为合格与不合格的计数值来进行统计分析，这时就属于计数值了。

1.3 常用数理统计方法

在生产与科学试验中，数理统计方法有广泛的用途，各种方法的原理及具体应用将在以后分章讨论，在此仅就几个主要方面作简要的说明

1. 参数估计

在生产和科学研究中，为了取得某种设备、工艺过程或产品质量的可靠技术经济指标，一般应做若干次重复试验(或观察)。但每次所得结果必然是参差不齐的，我们究竟应该怎样给出试验结果呢？通常只给出各次试验结果的平均值。这种做法比只凭一次试验的结果当然要准确得多，但还不完善。比如一台浓缩机的处理量，经多次反复测试平均为 $2.5 \text{米}^3/\text{米}^2 \cdot \text{时}$ ，我们并不能因此就断定它就是这台浓缩机真正的处理量，因为浓缩机的处理量受各种随机因素的干扰并不是一成不变的，有时会高于 $2.5 \text{米}^3/\text{米}^2 \cdot \text{时}$ ，有时也会比它低。比较完善的办法是给出一个区间估计，比如 $2.5 \pm 0.3 = 2.2 \sim 2.8 \text{米}^3/\text{米}^2 \cdot \text{时}$ ，并且给出这种估计的可靠性(把握)有多大，比如为95%。这就是说这台浓缩机的处理量不大可能超出 $2.2 \sim 2.8 \text{米}^3/\text{米}^2 \cdot \text{时}$ ，超过它的可能性很小，只有5%。用这种方法给出试验结果显然比只给出一个平均数要严密可靠得多。这里给出的区间称为“置信区间”，这个95%称为“置信概率”。类似地我们也可以对两种设备(或工艺)指标间的差值，或仪器的分析精度等其它指标作统计推断。

2. 统计检验

在技术革新中往往遇到要鉴定对比一种新方法、新工艺或新

设备的性能是否比原有者更有效，通常要在相同条件下作对比试验。只对比一次显然并不可靠，重复进行多次对比又会发现结果有好有坏，怎么下结论呢？一般也是对比平均值，比如新工艺的效率平均达到了95.4%，而原有工艺平均只有93.5%，如果不对试验资料进行统计分析，就据此断言新工艺一定比原有工艺更有效，那是很危险的。因为试验中不可避免要受到随机因素的干扰，新工艺的效率并不都能达到95.4%，原有工艺的效率也有比93.5%高的时候。两者效率平均值间的这种差异，究竟是两种工艺本身确有优劣之分，还是仅仅是由于随机因素干扰所造成的，尚难区分。正确的做法应该对试验资料进行统计分析，排除随机误差的干扰，从中作出有一定把握（比如按95%或99%）的结论来，这就是统计检验问题。统计检验的内容很多，在此不可能一一列举。

3. 可疑值的舍弃

科学试验中经常会遇到在一批相同条件下试验取得的数据中，有个别数据过大或过小，我们自然要怀疑这次结果是否正常。如果确实发现此时试验中某些环节、操作有失误，当然可据此将它舍弃，但有时并没有发现什么异常原因，这时就不能凭主观意断作出取舍，而必须进行统计分析。只有统计分析证明这种波动确属已超出随机因素干扰的范围，才允许将它视为反常值而予以舍去。这种方法叫做可疑值的统计判别法，在数据处理中经常用到。

类似的用途是全面质量管理中的质量管理图。在质量管理图中，标明了稳定生产条件下产品质量指标正常波动范围的界限。在此界限内波动属于随机因素的干扰，是不可避免的正常现象。一旦超出这个界限，就预示着生产环节中可能出现了某些反常现象，提醒操作管理人员及时查找、改正。如何制定这个界限，以及如何根据指标的波动情况判断生产是否正常，就要用到许多数理统计方法。

4. 分析影响事物变化的因素

为了改进生产，提高工艺效果，寻求最优的生产条件，首先要弄清楚究竟有哪些因素对试验结果有影响。但是影响因素往往很多，它们对试验结果的影响又往往是交织在一起的。如果对每一个因素以及各因素间的每一种可能的配合方案都一一进行试验，工作量显然太大，有时甚至是不可能的。数理统计中的试验设计（比如正交试验法）能帮助我们正确地安排试验，用较少的试验次数而又能从中得出必要的结论来。此外，在分析各因素对试验结果的影响时，还必须将归属于各个因素及因素间相互作用（称交互作用）的效果，与不可避免的随机因素的干扰（试验误差）分离开来，才能对各因素的影响作出可靠性的判断，这就需要运用数理统计中的“方差分析”方法。

5. 相关关系与数学模型

工作中我们常常发现，有些参数之间往往存在着相互依赖的某种关系，例如煤炭的灰分与比重间，灰分与发热量之间，精煤回收率与精煤灰分之间等等。这些参数间的定量关系不可能从理论上找出，但是在一定的范围内两者间确实有很密切的联系，甚至是一种简单的线性关系。然而各实验点又不可能完全落在一条直线上，能否把它当成线性关系来处理，以及如何根据实验资料给它配一条误差最小的直线（或曲线），这就是“回归与相关分析”中要研究的问题。按照相关分析，如果证明了两者间确实存在着显著性相关并给出了它们之间的经验方程，我们就可以根据一个变量（自变量）的取值去预报另一个变量（因变量）的可能结果，并给出这种预报精度的估计。如果影响因素不止一个，也可以根据实验资料通过分析为它们选配合适的数学表达式，这就是所谓的数学模型。数学模型在自动控制及最优化问题中占有十分重要的地位。

6. 抽样检验

生产中对原料、产品以及各中间环节的产品质量要进行经常性的检查。我们希望检验结果能如实地反映被检产品的质量。然

而在采样、制样及化验过程中不可避免也要受到各种随机因素的干扰，致使检验结果与实际质量间存在着一定的误差。为了把误差控制在规定的允许范围内，又不致使工作量太大，必须选择适当的采样方法，确定子样重量及子样份数，为制样与化验工作制定合适的操作规程。在这些工作中涉及到许多数理统计方法。

上面列举的六个方面只是工作中最常遇到的，它并不是数理统计的全部内容，只是为了说明对于科技工作者来说，数理统计是一种十分有用的工具。