常用数理统计方法

中国科学院数学研究所统计组编

科学出版社

常用数理统计方法

中国科学院数学研究所统计组编

科学出版社

1979

内 容 简 介

本书着重介绍方法，便于广大工人、科技人员在工作中掌握和应用，介绍时以具体例子为主．每种方法只介绍直观的原理，不作数学上的严格推导．书中选择的方法，是最常用的一些数理统计方法，如：数据整理、统计检验、正交设计、方差分析、回归分析、抽样检查等．为了使用的方便，书中还介绍了一些更简捷的方法，如非参数方法和概率纸．

附录部分收集了一些有成效的实例和常用的统计图表，供读者参考．

可供高中以上文化程度的工人、技术人员和科研工作者参考，亦可作为高等工业院校的教学参考书．

常用数理统计方法

中国科学院数学研究所统计组编

*

科学出版社出版

北京朝阳门内大街137号

天津市第一印刷厂印刷

新华书店北京发行所发行　各地新华书店经售

*

1975年10月第一版　开本：850×1168　1/32

1979年3月第三次印刷　印张：8 1/2　插页：1

印数：96,701—203,300　字数：220,000

统一书号：13031·114

本社书号：228·13—1

定价：0.85元

前　言

伟大领袖毛主席教导我们，从事一切工作都要**胸中有"数"。这是说，对情况和问题一定要注意到它们的数量方面，要有基本的数量的分析。**要做到这一点，必须深入现场，参加实践，得到第一手资料。得到资料仅仅是分析问题的开始，还要通过这些资料，进行分析研究，去伪存真，由表及里，抓住事物的主要矛盾。数理统计这门科学，可以帮助我们分析资料，处理数据；帮助我们科学地安排试验，制定抽样检查方案，为增加产品产量、提高质量提供了途径。

当前，我国社会主义革命和社会主义建设形势一片大好，新的形势对产品的数量和质量提出了新的要求。广大工人、科技人员迫切希望有一本通俗易懂的数理统计方法书籍。近年来，我们举办了几期数理统计学习班和讲座，这本书就是在原有讲义基础上加以整理修改而成。

本书着重介绍方法，以读后能够基本掌握和运用为准。介绍时以具体的例子为主，每种方法只介绍直观的原理，不作数学上的严格推导。具有高中文化程度的同志，就可以看懂。书中选择的方法，是最常用的一些数理统计方法。对于想知道数学原理的读者，或者解决更深一步的问题，我们在书末推荐了有关的参考书。还有一些常用的数理统计方法，有的已有中文专著（如质量评估），有的内容较深（如计量的抽样方案），本书就没有收集在内。

为了使用的方便，为了使文化程度不高的同志易于掌握，我们特地选择了一些更为简捷的方法，如非参数方法和概率纸的运用。对试验的分析，一般是用方差分析的方法，本书尝试用一种简单的直观分析法，使得掌握起来更为方便。附录中收集了一些有成效的实例，对读者有一定的参考价值。实例中有一部分用了较复杂

一点的方法，可作为进一步理解这些方法的参考.

第一章介绍了一些数据整理的方法以及数理统计中最基本的一些概念. 读完第一章，读者可按自己的需要选择阅读章节，不必拘泥于原来的顺序，因后面几章有一定的相对独立性. 凡是标有*的节，第一次可以不看.

我们的工作曾得到北京大学、青岛市科技组、北京市纺织局、北京维尼纶厂、北京国棉二厂、北京印染厂、青岛啤酒厂、青岛国棉八厂、青岛橡胶九厂、中国科学院各兄弟所等单位的领导和有关同志的大力帮助和热情支持，在此表示深切的感谢.

由于我们的水平很低，实践的经验不多，书中定有许多错误和缺点，请同志们批评指正.

编　者

1972 年 国庆

目　录

第一章 数据整理

§1 引 言

在生产斗争和科学实验中，经常要接触许多数据．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，是人们为进一步增加生产、提高质量而采取措施的依据．

但是，这些情报往往并非一目了然，而是蕴藏在大量数据之中．我们必须去粗取精，去伪存真，对数据作科学的整理和分析，尽可能充分和正确地从中提取出情报来．

本章介绍有关数据整理的基本方法，使我们了解数据为什么需要整理和如何整理；这一章还介绍了数理统计中经常用到的基本概念和术语，为以后各章的学习打下基础．

§2 基本概念

要整理数据，首先必须了解数据的属性．例如，在20天内，从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，有如下100个数据：

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

容易看出，这些数据有如下特性：

1. 波动. 在同样条件下生产出来的纤维纤度并不完全一样，表现出一定的波动.

2. 规律性. 数据虽有波动，但不是杂乱无章的，而是呈现出一定的规律. 仔细观察一下上列数据，就会发现它们都在 1.27 到 1.55 之间. 在 1.36 到 1.44 之间，机会多一些，在其它范围，机会少一些. 如果在同样的生产条件下再抽取一批数据，将发现纤度波动的情况和前一批数据十分相似，而且纤度落在任一范围的数目在该批数据中的比例是比较稳定的. 这个稳定的比例称做“概率”. 这就是说，纤度散布的情况是有规律性的.

象纤度这样的例子，在生产实践中是经常遇到的. 如炼铁厂每炉铁水的含碳量；某种产品的废品率；每年某月的降雨量；某种橡胶产品的磨耗量和强力等等，它们都有上述两个性质，即既有波动又有规律. 数理统计就是从有波动的数据中找出其规律性的一种数学方法.

为了今后叙述的方便，先引进几个概念：

1. 总体和个体. 我们所研究的对象的全体叫做总体，其中的一个单位则叫做个体. 譬如我们研究在正常生产条件下维尼纶的纤度，那么凡是正常生产条件下生产的纤维，其纤度的全体就是一个总体，而每一个纤度则是一个个体. 当研究的对象改变时，总体和个体也随之改变. 在整理数据之前，必须把它们弄清楚.

2. 样本. 总体的一部分叫做样本. 如上面列举的 100 个数据，就是从正常生产条件下的纤度这个总体中抽出的样本. 样本中所含个体的数目（如上例为 100），叫做样本的大小（或容量）.

一个总体所含个体的数目可以很多，甚至无穷，以致不可能一一加以考察. 如纤度这个总体，它的个体数目多得数不清. 有时候，数据的测定是破坏性的，如研究炮弹的杀伤半径，测量一个就要爆炸一个. 因此，尽管总体所含个体的数目不是很多，也不允许全部加以考察. 我们只能通过样本来了解总体. 统计方法就是解决如何从样本来研究总体的问题.

为了介绍数据整理的方法，我们将以纤度这批数据为例，说明整理的步骤，以及如何从样本来研究总体.

§3 几个重要的特征数

我们从总体抽了一个样本，得到一批数据 $x_1, x_2, \cdots, x_n$.

工程技术人员处理这批数据时，经常用算术平均数来代表这个总体的平均水平. 这个方法简单而有效，统计中称这个算术平均值为“样本均值”，并记为 $\bar{X}$:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1.1}$$

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

有时候，为了减少计算，把数据 $x_1, x_2, \cdots, x_n$ 按大小次序排列，用排在正中间的一个数表示总体的平均水平，称它为中位数. 当 n 为奇数时，正中间的数只有一个；当 n 为偶数时，正中间的数有两个. 在后一种情形，中位数等于这两个数的算术平均.

但是，只反映平均水平经常是不够的，例如纤度太大、太小都不好. 即使平均水平符合要求，若数据波动太大，这批产品的质量还是不能令人满意的. 因此，数据波动的大小也是一个重要的指标. 如何度量波动的大小呢？一种简单的方法是用极差. 极差是指数据中最大与最小之差，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 $x_1, x_2, \cdots, x_n$ 中最大的和最小的. 由于极差没有充分利用数据提供的情报，因此反映实际情况的精确度较差，于是人们想出了另一个测度——标准离差，或简称标准差

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2}. \tag{1.2}$$

也可用它的平方——方差 s^2 来衡量数据的波动. s 越大，波动越大；s 越小，波动越小. s 比极差反映问题精确，但计算比极差复杂得多. 各有优缺点，看具体情况而加以运用.

我们都有这样的体会，测量较大的东西，绝对误差一般较大；测量较小的东西，绝对误差一般较小. 因 R 和 s 只反映绝对波动的大小，因此还应考虑相对波动的大小. 这在统计上用变异系数

$$CV = \frac{s}{\bar{X}}$$

来表达. 在纺织系统中人们称变异系数为不匀率.

在很多情况下，只了解数据的几个特征数还是不够的，例如在制定产品的抽样方案时，需要了解波动的更完整的规律. 波动的规律不同，抽样方案也就不同，于是从实际问题中又抽象出分布的概念.

§4 频数分布和频数分布函数

要弄清数据波动更完整的规律，必须找出频数分布. 为此要将数据分组，我们以纤度的数据为例，说明分组的步骤.

1. 找出最大值与最小值. 在第 2 节所列纤度数据中，最大值为 1.55，最小值为 1.27.

2. 决定组距和组数. 在样本比较多时，通常分成 10～20 组. 样本数少于 50 时，分成 5～6 组. 先决定组距，然后定组数. 组距决定于极差 R. 在本例中 $R = 1.55 - 1.27 = 0.28$. 看来组距定为 0.03 比较好，这时共分 10 组. 需要说明的是，并不是在所有情况下都采用等距分组，要具体情况具体分析.

3. 决定分点. 如果我们按纤度原来的测量精度分组，即分为 1.26～1.29，1.29～1.32，…，那么对纤度恰好是 1.29 的样本，是分到 1.26～1.29 这一组呢？还是分到 1.29～1.32 这一组呢？为了避免这种麻烦，通常要使分点比原测量精度高一位. 于是分成如下 10 组：

1.265～1.295，1.295～1.325，1.325～1.355，…，1.535～1.565.

4. 数出频数. 用选举唱票的办法数出样本落在每个组的数目，称为频数. 唱票时把它列成表 1.1 的形状. 这样得到的分组频数表，称为频数分布表. 频数与样本总数之比称为相对频数.

表 1.1 频数分布表

分　　组	频　数　计　算	频　数	相对频数
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		$n=100$	1.00

在求出了频数分布表以后，能够比较清楚地看出数据波动的规律. 为了更加直观起见，把它划成直方图. 在横坐标上标出分组的点，纵坐标对应频数，以组距为底边划出高度为频数的矩形，便得图 1.1. 这种图在统计上叫做直方图. histogram

可以设想，如果我们取更多的样本，组分得更细，那么，各样本值或者各组的相对频数将趋于一个稳定的值. 因此，纵坐标如果不取频数而取相对频数，就得到相对频数分布直方图(图 1.2). 这时，直方图的形状逐渐趋于一条曲线. 换句话说，作为相对频数分布的极限，可以考虑一个

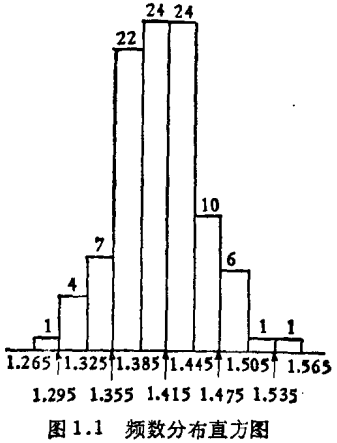

图 1.1 频数分布直方图

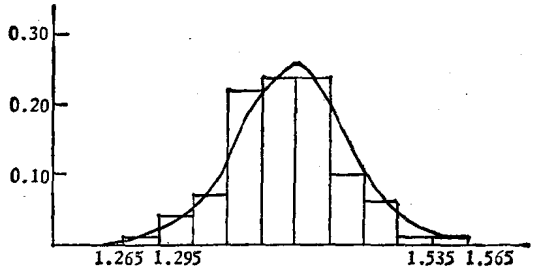

图 1.2　相对频数分布直方图

稳定的函数. 如本例那样，当样本值是连续变量的情形，这个函数将表达一条光滑的曲线.

这条曲线排除了抽样和测量的误差，完全反映了纤度波动的规律. 这种曲线在统计中很重要，叫做“频数分布曲线”. 若数据波动的规律不同，分布曲线的形状也就不一样. 在实际中，形如图 1.2 的曲线最多，应用也最广，称为正态分布曲线.

由于相对频数之和等于 1，不难看出，如果纵坐标取为

相对频数/组距，

那末直方图各矩形面积的总和等于 1. 换句话说，分布曲线与横坐标所夹的面积等于 1.

本书将着重介绍正态分布，对其它分布只作概括的介绍.

§5 正 态 分 布

正态分布的形状如图 1.2 所示，曲线有个最高点. 以此点的横坐标为中心，对称地向两边快速单调下降. 遵从正态分布的例子很多，如炼铁厂每炉铁水的含碳量；同一型号的铆钉头的直径；纤维的强力；健康人红血球的数目；随机测量误差等等.

正态分布曲线由正态概率密度函数

$$\varphi(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.3}$$

给出，式中 x 是从此分布抽出的随机样本值；$e\doteq 2.718$ 是自然对

数的底；μ是曲线最高点的横坐标，叫做正态分布的均值，曲线对μ对称；σ的大小表达曲线胖瘦的程度，叫做正态分布的标准离差. σ越大，曲线越胖，数据越分散；σ越小，曲线越瘦，数据越集中，如图 1.3 所示.

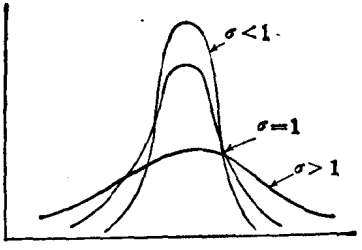

图 1.3 σ的直观意义

可见，有了均值μ和标准离差σ，就可以把正态分布曲线完全确定出来. 当$\mu=0$，$\sigma=1$时的正态分布，叫做标准正态分布. 为了叙述的方便，今后用$N(\mu,\sigma^2)$表示均值为μ、标准离差为σ的正态分布，于是$N(0,1)$就表示标准正态分布.

纤度的例子，从直方图形状来看，很象正态分布. 如何来检验我们的猜测呢？通常有偏态峰态检验法，χ^2检验法. 然而这两种方法的计算比较麻烦，可参考[1]. 这里介绍一种简单而有效的判别法则. 有一种特殊的坐标纸，叫做正态概率纸，它的横坐标是普通的刻度，纵坐标按正态分布的规律刻划. 按照规定的方法在正态概率纸上打点，如果分布是正态的，则所打的点几乎在一条直线上. 正态概率纸的样子，见附录 II 图 1. 现在我们就用正态概率纸来检验纤度波动的规律是否为正态分布.

表 1.2 累计频率表

分组右端点	频 数	累计频数	累计频率(%)
1.295	1	1	1
1.325	4	5	5
1.355	7	12	12
1.385	22	34	34
1.415	24	58	58
1.445	24	82	82
1.475	10	92	92
1.505	6	98	98
1.535	1	99	99
1.565	1	100	100

首先，在表1.1的基础上列出表 1.2. 表的第三列累计频数是到该组为止的频数和，第四列累计频率是以样本总数 100 除第三列的数，以%表示. 在正态概率纸上以表 1.2 的分组右端点为横坐标，累计频率为纵坐标. 例

如,第一个点的横坐标是 1.295,纵坐标是 1;第四个点的横坐标是 1.385,纵坐标是 34,其余类推. 打点结果,如图 1.4 所示. 从图上看出,这些点近于一条直线. 照顾到各个点,可用直尺划一条直线把它们连结起来. 如果数据遵从正态分布,那么这些点应该在一

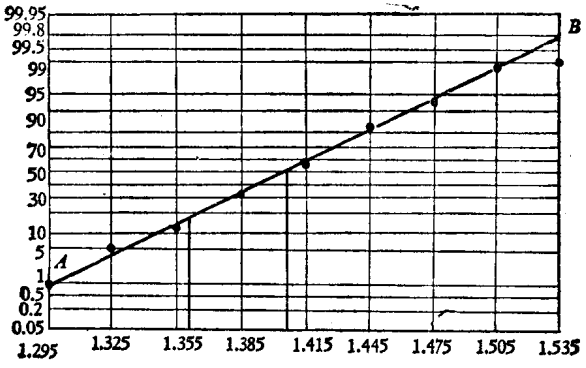

图 1.4 正态概率纸

条直线上. 但是由于样本的随机波动,多少有点偏差,这是允许的,但偏差不能过大. 过大了,我们就怀疑总体不是正态分布. 一般地说,中间的点离直线的偏差不能过大,两头的点偏差可以允许大一些. 对纤度的例子来说,从图上看,除最后一点外,其余各点离直线的偏差都不太大,故可认为纤度是遵从正态分布的.

只知道纤度是正态分布,对解决实际问题,是不够的,还需要知道它的均值 μ 是多少,标准离差 σ 有多大.

从图 1.4 可以近似地估计出 μ 和 σ. 以后 μ 和 σ 的估计值分别记作 $\hat{\mu}$ 和 $\hat{\sigma}$. 由直线 AB 与纵坐标 50% 的交点向下作一条垂线,交横轴之点即为均值. 从图上看出,均值的估计值约为 $\hat{\mu}=1.405$. 直线 AB 与 15.9% 交点的横坐标是 $\mu-\sigma$. 从图上看,这个值约为 1.358. 于是标准离差的估计值 $\hat{\sigma}=1.405-1.358=0.047$. 只要图画得准,这些估计值还是比较精确的. 有关正态概率纸的用法,可参看[2].

知道了纤度遵从正态分布,给指导生产有很大好处. 对于