

近代数学讲座讲稿

- (一) 数理统计简介
- (二) 相关回归分析大意
- (四) 略谈运筹学

北京教师进修学院
1960年

近代数学讲座讲稿(一)

数理统计简介

科技大学 周云龙

北京教师进修学院印发 1960.9

Ⅰ、一提到数理统计这一学科，不少人有这种看法，认为数理统计仅是收集大量数据资料，以及将这些用图表表示出来，在数学上说最多不过是计算一下平均数，百分比等这类工作，其实这种看法是三四十年前极为陈旧的观点。上面提到的那些多少有些机械性质的工作，仅是现代数理统计中很次要的一部份。

数理统计可以说是概率论的源泉，也是概率论应用的对象，它应用概率论中的理论，通过整理和分析由观察或实验得出来的大量数据，来研究大量偶然现象所呈现的规律性，它所用的基本方法就是抽样的方法，即“从了解到部分的情况来推断全体情况”。而这种推断其结果的正确性是用概率形式来表达，也只有这样，才能使我们对于推断其结论的正确性有正确的了解。在实际应用数理统计时，大致可分为三步：

- 一、进行试验，或调查研究收集材料；
- 二、根据试验所得资料，作主观的结论；
- 三、将这些结论表示成概括的定律，在有些问题中尚须进行一些验证。

Ⅱ、~~数理统计~~ ~~这门~~ ~~学科~~ ~~的~~ ~~内容~~ ~~非常~~ ~~广泛~~，在此只能提些常
~~见的~~ ~~例子~~ ~~来说~~ ~~明~~ ~~它~~ ~~的~~ ~~意义~~，并再通~~过~~具体的例题来介绍问题的类型。
~~在~~ ~~工~~ ~~业~~ ~~生~~ ~~产~~ ~~中~~ ~~生~~ ~~产~~ ~~了~~ ~~一~~ ~~大~~ ~~批~~ ~~某~~ ~~一~~ ~~品~~ ~~种~~ ~~的~~ ~~电~~ ~~子~~ ~~管~~，而需

~ 2 ~

了解这批电子管的质^性量情况。其废品率是多大，这里规定电子管^性使用小时数达到某一预定数值认为合格，不然认为是废品。为了了解质量，今从中抽出 n 个加以试验，这 n 个的使用时间分别为：

$x_1, x_2, \dots, x_n$ 。

现在就从 n 个电子管的质^性量情况来推断这批电子管的质^性量情况。

很显然，我们立即有一系列的问题：怎样才能保证抽出的这 n 个最有代表^性呢？抽多少个最合适呢？（即 $n=?$ ），容易看到，如果 n 太大，人力物力耗费太大， n 太小，得出的结果又不见得可靠。另外，用一部分的质量情况来推断，不可能是百分之百可靠，因此谈到可靠性，只能通过概率的形式表现出来。

上面问题的特点是，有了一个我们所研究的全部对象组成的总体（上例中这个总体是全部电子管）这个总体的某个指标是我们密切关心的（如上例中的质量指标废品率），而我们从这个总体中抽出若干样品的研究来推断我们所关心的指标的数值。

象这一类问题有很多，例如，如果上面这些电子管的质量指标是平均使用小时数，那么我们可用抽出 n 个的平均使用小时数来推断。又如，为了想知道某一森林的木材储藏量，显然不可能把这森林全部砍伐后看个究竟，也是抽出森林若干部分来研究，从而估计整个森林的储量。

上面所谈的这一类问题，即通过我们所要研究总体的一部分——样本，来估计总体某个或某些指标的问题，这就是统计中的估计问题，或称参数估计问题。

2. 数理统计第一类重要的问题，可以通过下面的例子来说明。

为了加快某种化学反应的速度，在我们可利用的两种催化

剂中，确定那一种好，在事先我们往往不知道那种较好，还是
没有差别，这就通过试验来加以解决，比方说做了若干次试验
后，实验资料表明用催化剂甲的速度 V_1 ，比用催化剂乙的速度
 V_2 大一些，但我们不能马上说甲一定比乙好，因为我们的根
据若干次试验资料来看 V_1 确实比 V_2 大一些，实际上有可能
甲、乙差不多，甚至甲可能比乙小些，只是由于在试验中受到
一些偶然因素的影响，使 V_1 比 V_2 大出了一些，如果 V_1 比 V_2
大很多，那末把这种差别归之偶然因素的影响的根据就很小，
从而推断甲乙间的差别，由于甲比乙好而不是偶然因素的影响
的根据就大了，但问题是甲要比乙大多少才能认为不受偶然因
素的影响？又试验的次数应该是多少次呢？

这一类型的问题属于数理统计中的“假设检验”的范围，
解决这类问题常是这样：开始我们先立一个假设“催化剂甲
、乙没有差别”然后根据所作试验而得的材料加以分析，然后
根据分析的结果来决定采纳这假设还是否定假设。

例如：甲方提供一批产品给乙方，而乙方要求该批产品的
废品率不能超过 0.05，在不可能把全部产品逐一检查情况下，
必须从抽出一部分的废品率来推断全部的情况以决定接受还是
拒绝。

3、我们知道，农作物的产量的高低，主要因素之一是
施肥量的多少，它们之间有密切的关系，如果把影响产量的
其他因素，如种子品种、年平均温度、年平均雨量等等，都
全部加以固定（显然，实际上是办不到的），那么，还是会发现
施肥量确定后，还不能唯一地决定产量，因为有不少随机
性的，为人们无法控制的因素在起作用，因此虽然都施同样的
肥量，产量还可以不同，肥量与产量间的关系并不是我们
熟悉的函数关系，很显然，最合理最方便的看法是把产量看

域一定与施肥量有关的随机变量。否则在施肥量又决定时，不能唯一地确定 y 的数值。但能确定 y 数值的分布情况，也就是取得某些值的概率情况，这样的关系，我们称为相关关系。

相关关系不能通过一定的函数形式表达出来。在应用上当然不方便，因此我们企图找一种函数关系，近似地代替所研究的相类关系，这就涉及到所谓回归问题。

4. 我们常常会碰到为了要比较不同品种种子的优劣，几种不同肥量的效果，或者在工业中比较不同生产操作方法的好坏，就要进行试验。然而要怎样进行试验才能最好地达到我们的目的？怎样安排试验才能排除非随机因素的影响？如何使试验成本最低？这类问题就归之于“试验设计”中了。

四. 数理统计中的基本术语：

由于和各方面问题有密切联系，这些术语是非常直观的，但是具体问题还要具体理解，不能一概而论。

1. 总体（母体）：它就是指我们所研究的全部同类对象组成的一个集体，我们不能把总体看成在质的方面毫无联系的多元素（组成总体的个体称为元素）的一个偶然堆积。总体中的元素不是毫无联系的，必须基本上是同质的，但它们之间的联系却也不单他是必然联系的。因为如果是必然的联系的话，知道第一个元素能完全肯定第二个元素的性质如何，知道第二个能肯定第三个如何，这样我们就不必对总体进行统计研究了。

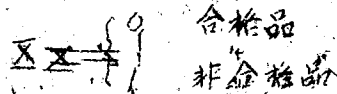
例如：我们考察某工厂在某天生产电子管的平均使用时间，那么我们的总体就是该厂这天生产的全部电子管所组成的集体，而我们研究的对象是电子管的好坏，所以电子管是组成总体的元素，很明显的看到这批产品中各个电子管在质的方面不见得会毫无联系，因为这是同一工厂同一天生产的。而这工厂

的机件工作状态处在较稳定的状态下，那么每个电子管使用的
时间不会差别很大，即使有差别，但个数的也不很多，但是在检
查时从中抽几个试验出使用时间后，我们绝对不可能推断抽
第二个使用时间必是更多，点了二个也不可能推断第三个必
果多少，这表示这些电子管间没有单纯的必然联系。

例如：一辆汽车的许多不同种零件不能看成总体，因为这
些零件不是同类的元素，但是，如果我们要对一辆汽车上用一
标准的螺栓的某一指标（如长度）进行考察的话，它们倒可以
看成一个总体。

有时根据我们所考察的目的确定总体。例如，考察某棉色
纤维平均长度时，那么由棉花纤维的集合构成一个总体，若考
察棉色平均重量时，棉色为元素构成的集体是总体。

2. 总体的某指标及其分布：这是一个十分重要的概念，简
单地说：总体的指标是为了一定目的而定义在元素上的数值，
例如，总体是一批电子管时，为了了解质量，其指标是电子管
使用时间，那么对于每个电子管这指标（使用时间）有唯一的
一个数值与之对应，当然在有些情况下，总体的指标只是性质
上的东西。例如，有考察某批产品的废品率，对每个产品（元
素）来说只是“合格”、“不合格”这一个性质，但是我们能
在这总体上定义一个函数。



这成了前一情况

实际上，总体的指标是个随机变量，比如影响电子管使用
时间的主要因素之一，是阴极放射电子的能力，在制造时由于
一些不能由人们控制的随机因素的作用，影响到阴极的质量，
从而影响到使用时间，因此使用时间是一个随机的数值，不是

~ 8 ~

人们事先能予料的。

因为这样的指标可以看成是个随机变量，那么可以象“随机变量的分布”一样引进“指标在总体内的分布”这一概念了。但是在实际上各种指标的概率分布是不知道的。在有些问题中就需要假设一个概率分布，近似地反映未知而客观存在的总体的分布，这个分布通常称为理论分布。但是这个理论分布，不是凭人类的主观随便假设，而是由于人类从经验中观察到某些具体问题时总结出来的。有了这种分布后，我们可以计算总体的平均值、均方差，这些都是在数理统计问题中最为我们关心的。例如，总体 Ω 是一批电子管，指标 X 表示 Ω 内每一个电子管使用时间， X 是随机变量，因此可以引进“ X 在 Ω 内的分布”，从理论上说可以知道“使用时间小于 x 的概率” $F(x)$ 了，我们是不知道分布的。但人类多年来的经验给出

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

其中 μ 是平均值， σ 是均方差。（这就是所谓的正态分布）

3. 样本：样本是按照一定的方式从总体中抽出的若干个元素组成的一一个集合，若样本内包含的元素个数是 n ，称 n 为样本大小，因为我们收集的是总体的某指标，而这指标常是用数字来反映，因此抽出大小为 n 的样本后就得到 n 个数字。

$$x_1, x_2, \dots, x_n \quad (1)$$

所以常称这些被抽出的元素的指标数值为样本，而不指样本元素的实物本身。

得到了样本后，可以计算它的平均值和方差

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

因为样本是随机抽的，所以，(1)中每个数值 X_i 都是随机变量，所以与(1)有关的 $\bar{X}$ 、 S^2 也随每次抽样的结果而不同，因此 $\bar{X}$ 、 S^2 也是随机变量，这和总体指标的平均值、均方差不同，总体的平均值均方差是因总体的分布不知道而不知具体数值，但在任何具体问题中它的数值是一常数。

样本是“按一定方式”从总体中抽得的，就表示抽取样本可以用不同的方法、怎样去抽样，抽出样本的代表性如何等？问题是“抽样技术”和“抽样方法”而专门研究的。下面我们介绍一下抽样几种方式而不加以详细讨论比较。

(a) 在总体中第一次随机抽出的元素将其指标值记下，然后放回总体，再随机地抽第二次，这样进行 n 次，所以得样本 $X_1, X_2, \dots, X_n$ 称为简单随机样本。

(b) 在总体中逐次随机抽出，但每次都不放回总体，这样所得的样本称为随机样本。

在工业技术中常用的是上述方法，在总体的元素个数很大时，这二种方法基本上没有差别，在元素个数很少时，就必须加以区别。此外尚有分层抽样、机械抽样、集团抽样。

IV. 统计资料的整理

现在我们假定通过试验或抽样得到了大批资料，这种原始资料往往是比较杂乱看不出规律，所以必须加以整理，整理的方法有二类，列表法与图示法。

(1) 列表法

我们把抽得的样本，依样本中元素指标数值的大小重新排列（若观察的问题与样本元素指标值次序无关的话），

例如，对布匹进行拉力试验 25 次，出现次序为（单位公

~ 8 ~

(4)

22, 24, 20, 23, 21, 19, 23, 22, 20, 22, 20, 22, 23,
25, 21, 21, 22, 24, 23, 22, 23, 21, 22, 21, 23;

表 1

重排次序得:

19, 20, 20, 20, 21, 21, 21, 21, 21, 22, 22, 22, 22,
22, 22, 22, 23, 23, 23, 23, 23, 23, 24, 24, 25

表 2

上表重排列形式

强 力	Σ	19, 20, 21, 22, 23, 24, 25
出现频数	f	3 5 7 6 2 1
出现频率	f/n	$1/25$ $3/25$ $5/25$ $7/25$ $6/25$ $2/25$ $1/25$

表 3

这样, 对所抽样本反映的情况从表 3 就一目了然了; 有时从抽得样本后可直接制成表 3。

有时样本很大, 例如, 观察 200 个零件的长度与标准长度的偏差 (单位厘米, 资料不全写出)

1.0, 1.5, 7.5, ..., 30.5, -1.5, -2.0, ..., -8。

对很大样本常分组处理, 大小为 200~300 的时候, 常分成 10~20 组。(若总体是正态分布常分成 12 组)

今将样本中出现最大值 30.5, 最小值 -20 取来, 则差距 $30.5 - (-20) = 50.5$, 今分成 10 组, 故每组间距为:

$$\frac{50.5}{10} = 5.05 \approx 5. \quad \text{今再制出下表}$$

组号	组的界限	频数 f_i	频率	组的中点
1	-20 — -15	7	0.035	-17.5
2	-15 — -10	11	0.055	-12.5
3	-10 — -5	15	0.075	-7.5
4	-5 — 0	24	0.12	-2.5
5	0 — 5	49	0.245	2.5
6	5 — 10	41	0.205	7.5
7	10 — 15	26	0.130	12.5
8	15 — 20	17	0.085	17.5
9	20 — 25	17	0.085	22.5
10	25 — 30	3	0.015	27.5

表 4

将表说明一下，第一组的距离为 -20 ~ -15，今在抽出样本的 200 个测量中，发现发生偏差在 -20 到 -15 之间为 7 个，因此频数为 7，而频率为 $7/200 = 0.035$ ，组中点 -17.5 是 -20, -15 二数的中点，我们在样本很大时作成表 4，不但比较明显地反映出而且在将来计算平均值和方差时带来不少方便。

(2) 画法:

(a) 在横坐标轴上标出各组间隔的端点，以及每组的组中点，并组中点作一直线垂直横轴，其高度为组中频数，然后将各线段的端点连成一折线，即成多边形图形（见图 1），此处三轴采用不同单位，不然图形将太靠近横轴）。

(b) 在横坐标每组区间上作一长方形，使面积等于每组内的频数，标出的图形称为直方图。（见图 2）

这里二种作图的方法都与分组有关。

~10~

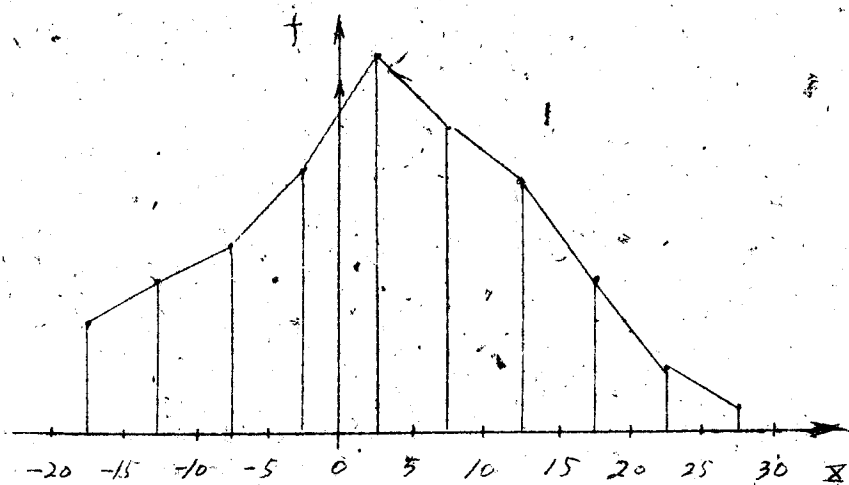


图 1

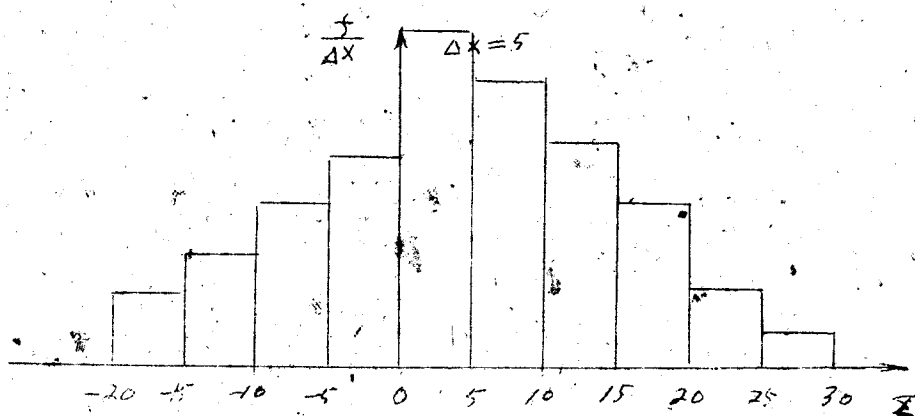


图 2

样本均值方差的计算，

在表 1 中所示样本平均值为

$$\bar{x} = \frac{1}{25}(22+24+20+\dots+22+21+23) = 21.96$$

$$S^2 = \frac{1}{25}[(22-21.96)^2 + (24-21.96)^2 + \dots + (23-21.96)^2] = 1.96$$

有时因样本太大，直接计算比较困难，为以简化。

(a) 计算平均值查点法。

此法要点是将各数都减去同一数 C (数 C 任意选取) 将其剩余数再计算算术平均值后再加上 C 便得原来算术平均值，即

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n (x_i - C + C) = \frac{1}{n} \sum_{i=1}^n [(x_i - C) + C] \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - C) + C\end{aligned}$$

(b) 计算均方差的查点法。

由公式

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

由于 $\bar{x} = \frac{1}{n} \sum_{i=1}^n (x_i - C) + C$ 以及

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i - C + C - \bar{x})^2 = \\ &= \sum_{i=1}^n [(x_i - C)^2 + 2(x_i - C)(C - \bar{x}) + (C - \bar{x})^2] \\ &= \sum_{i=1}^n (x_i - C)^2 + 2(C - \bar{x}) \sum_{i=1}^n (x_i - C) + n(C - \bar{x})^2 \\ &= \sum_{i=1}^n (x_i - C)^2 + 2n(C - \bar{x})(\bar{x} - C) + n(C - \bar{x})^2 \\ &= \sum_{i=1}^n (x_i - C)^2 - n(\bar{x} - C)^2\end{aligned}$$

因此，代入 S^2 公式得

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - C)^2 - (\bar{x} - C)^2$$

~12~

特别当 $C=0$ 时,

$$S^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2$$

(c) 分组时计算样本的平均值及均方差.

假定大小为 n 的样本已经分成了 k 组, 组距为 C , 组中值分别为 $a_1, a_2, \dots, a_k$. 每组的频数分别为 $f_1, f_2, \dots, f_k$. 再把第一组的 f_1 个数值都看成是 a_1 , 第二组 f_2 个都看成 a_2 , 等等. 这样一来样本的总数没有改变仍是 n , 但每组的数值有些改变, 不是原来直接所得, 但是在 n 很大时这样做误差将大大简化, 这时,

$$\bar{X} = \frac{\sum_{i=1}^k f_i a_i}{n}$$
$$S^2 = \frac{\sum_{i=1}^k f_i (a_i - \bar{X})^2}{n}$$

我们为简化计算, 把 $\bar{X}$, S^2 的公式用另一种形式表达, 今在分组中值 $a_1, a_2, \dots, a_k$, 任意取定一个用 A 来记.

令 $u_i = \frac{a_i - A}{C}$ (注意 u_i 现在取整数值) 这样

$$\begin{aligned} \bar{X} &= \frac{\sum_{i=1}^k f_i X_i}{n} = \frac{C \sum_{i=1}^k f_i (X_i - A + A)}{n} \\ &= \frac{C}{n} \left[\sum_{i=1}^k \frac{f_i (X_i - A)}{C} + \sum_{i=1}^k \frac{f_i A}{C} \right] \\ &= \frac{C}{n} \sum_{i=1}^k f_i u_i + A \end{aligned}$$

这是因为 $\sum_{i=1}^k f_i = n$ 所以

$$\bar{x} = C \frac{\sum_{i=1}^k f_i u_i}{n} + A$$

经过计算可得

$$S^2 = C^2 \left[\frac{\sum_{i=1}^k f_i u_i^2}{n} - \left(\frac{\sum_{i=1}^k f_i u_i}{n} \right)^2 \right]$$

这二公式再结合表格一起可以很快计算出 $\bar{x}$, S^2 , 如, 以表 4 为例, 取第 5 组之中值 2.5 为 A , 此时 $C=5$

组号	组中值 a_i	频数 f_i	u_i	$f_i u_i$	$f_i u_i^2$
1	-17.5	7	-4	-28	112
2	-12.5	11	-3	-33	99
3	-7.5	15	-2	-30	60
4	-2.5	24	-1	-24	24
5	2.5	49	0	0	0
6	7.5	41	1	41	41
7	12.5	26	2	52	104
8	17.5	17	3	51	153
9	22.5	7	4	28	112
10	27.5	3	5	15	75
合计		$\sum f_i = 200$	$\sum u_i = 5$	$\sum f_i u_i = 72$	$\sum f_i u_i^2 = 780$

因此

$$\bar{x} = 5 \cdot \frac{72}{200} + 2.5 = 4.3$$

$$S^2 = 5^2 \left[\frac{780}{200} - \left(\frac{72}{200} \right)^2 \right] = 94.26$$

