高等数学第二部分

数理統計

高等数学教研組編

西安交通大學

1962.4

目 41606

数理統計

緒論

数理統計学和概率論一样是研究大数現象在数量上的規律性的科学，但是它却使用着不同的方法．較明確地說，数理統計是从实际資料出发来研究大量現象的規律性．

显然，如果我們进行了大量的观察，那就能看出这些大量現象所呈現的規律性，但是客观上又只允許我們对大量現象进行局部的观测．从表面看来这是矛盾的，然而我們只要能充分利用观察資料和局部与整体之間的內在联系来进行分析与推断，仍然能够認識这些規律性．因为这种从局部观察去推断整体的方法有着普遍的意义，因而数理統計的应用非常广泛．

数理統計方法在工业生产及科学研究中有着广泛的应用．下面举几个例子来說明数理統計所能解决的某些問題．

1．在現代大批生产的情况下，特別是工业自动化的发展，使产量不断提高．在这种情况下如何檢查生产出来的大量产品的質量成为严重問題．比如說一台自动机床日产1500个另件．出于对工作和經济的考虑，我們不能把另件每一个都进行檢查，而是从中間取一部分出来进行檢查，从而断定一天产品的質量情况．尤其是有很多試驗具有破坏性的，不允許我們对产品进行全面檢查．比如說檢查电灯泡的能点时数，要把电灯泡一直点到不亮为止．要檢查砲彈的射程必需发射一下，那末砲彈就炸掉了．諸如此类的例子在工业生产中很多．此外，还要由抽查的一部分产品質量情况，应决定整批产品合格与否，这就是产品驗收問題．

2．上面講的产品質量檢查与产品驗收是在一批产品已經生产出来，然后进行質量檢查，决定产品合格与否．这还是消极的．积极的方法是在生产过程中随时地进行一些檢验，来判断生产的进行是否正常，亦就是說生产出来的产品是否合于規格要求．这可启示我們随时調整机器使生产經常保持正常状态．这是所謂生产中的質量控制問題．

3．在生产中常要作这样的試驗：比較运用好几种原料（或者不同方法）生产出来的产品質量是否相同，进而判断那些原料（或生产方法）生产出来的产品質量較高．这时我們只能用各种原料（或各种生产方法）做

不多几次試驗，从試驗資料来作出决定。在科学研究中，常要在各种不同条件下做不多几次試驗，从試驗資料判断由于条件的不同对于試驗結果是否有显著影响。若不同，进而可以断定在怎样的条件下試驗得到的結果最好。这是所謂試驗分析問題。

4. 在生产中常需要找出所用原料与生产出来产品的質量之間的关系。在科学研究中要寻找各种因素与試驗結果之間的关系。比如說，在紡織厂中要找出原棉的物理性能（有纖維强力，纖維支数等）与成紗的質量之間的关系，以便如何合理地利用原料紡出高質量的紗。

在上面一些例子中都是从做一些試驗或作一些观察得到的一些数据来判断总的情况。这正是数理統計的任务，亦是数理統計所用的方法。

在数理統計中要解决二个問題：1.如何来設置观察或試驗。2.如何通过对观察或試驗得到的数据进行分析，科学地判断总的情况；更具体地說揭示大量現象中事物的本質（如1、2、3）及存在于一些事物之間本質的联系（如4）。总之，数理統計是通过对某些观察或試驗結果的分析来揭示存在在大量現象中事物本質和事物間的本質联系的科学。

数理統計是一門数学，它与各門具体学科有着密切联系。把它的方法与原理运用去研究各門科学，得到各种密切联系这門学科实际的数理統計学。如：工业应用数理統計，紡織工业数理統計，水文学中数理統計，統計物理、测量誤差理論等等。数理統計这門数学的理論与方法由于生产实踐和自然科学研究的刺激得到不断发展，反过来利用新的数理統計的方法能更广泛地去解决生产实踐和自然科学研究中提出的一些問題。

在运用数理統計方法时，首先要求从观察与試驗得到的数据具有高度正确性，对这样資料进行統計分析得到的結論才有效。再則統計分析所得結論是否正确，还要在实踐中进行檢查。

数理統計与概率論有什么联系呢？可以这样說，数理統計是以概率論作为工具的。更明确地說，在講解統計原理与导出統計方法时要用到概率論中的一些概念（如：概率，随机变数等）和一些結論（如：大数定律，中心极限定理）。

在本講义中将介紹一般的統計原理和方法。至于有些断言和式子的导出相当复杂，本講义只列出結論而不作推导。我們热忱期望工业高等学校学生在掌握了数理統計原理与方法之后能够运用去解决生产和科学研究中所提出的实际問題。

第一章 基本概念

在开始这一章中我們将介紹数理統計中的一些术語，討論概率論中的一些概念（如：隨机变数，分布函数，分布密度）与数理統計的概念如何結合起来，并运用概率論理論去解决統計中提出的某些最基本的問題.

§1. 总体，个体，样本．数理統計的目的和方法

所欲研究对象的全体称为总体（或母体）．这些研究对象具有通性．**总体中包含的基本元素称为个体．**

例 1. 檢查某天一机床出的产品的質量，那末这天这机床所产生出产品构成一个总体，而每一个产品是一个个体．总体中元素的通性是同一天同一机器同一工人生产出来的．

例 2. 檢查一包棉花纖維的强力与支数，这包棉花所有的棉纖維就构成总体，而每根棉纖維是一个个体．总体中元素的通性是同一包棉花中的纖維．

但是通常总体中的个体是指物体某种性質数量特征，并非指物体本身．比如說在例 1 中生产出的是圓柱形另件，那末所有另件的直徑构成一总体，所有另件的长度亦构成一总体，所有另件数重量亦构成一总体（如果需要考虑重量的話）．在例 2 中所有纖維的强力构成一总体，所有纖維的长度亦构成一总体．也有些物体的性質不可度量，这时可以任意給物体的性質一种数量标誌．如：若考察一批产品的質量，每一个产品仅有廢品和正品之分，我們可以假設（完全是人为的）廢品是"1"，正品是"0"．这批产品質量的总体可以看成由很多个"1"和"0"組成．

总体中所含个体的个数，称为**总体容量**，用N表示．总体容量可大可小．比如說我們考察一机器某天生产出 50 个产品的长度，那末总体容量是 50．又如一大包棉花中所包含纖維达 5000 亿根以上，这时纖維强度构成的总体的容量可以看成是无限大．而无限大总体亦是存在的，比如說一条河流的年流量，从它形成的一天起每年都有，这河流所有年流量构成的总体在理論上可以認为是无限的．**在数理統計中不妨把容量相当大的总体看成是容量为无限的总体．**这样对我們在数学上討論很有好处，在下面 § 4 中将加以說明．

个体既是指物体的某种性質的数量特征，在数学上只是考虑这一些質

的变量，今后一个个体用一个数来表示。那末总体就成为一些数的集合，用 $x_1, x_2, \cdots, x_N$ 或 $x_1, x_2, \cdots$ 表示。

我們再进一步分析通常所見的总体是怎样一些数的集合。我們发现，这一般可以分为离散和連續二种。如果我們一批（n 个）另件的直徑，各为 $x_1, x_2, \cdots, x_N$。則 $\{x_1, x_2, \cdots, x_N\}$ 是一个离散数集。如果我們測量一物体的长度，由于各种偶然因素影响，会引起誤差，每一次測量可能取得的数值可以充滿某一区間。这一区間是一个連續数集。应該特別注意，不管是总体是离散数集或連續数集，其中的数据都是按比例分布的。这点下节要詳細分析討論。

由于总体的容量一般来說都很大：因而要得到总体的性質經常是从总体中取出一部分加以研究。总体的一部分个体組成的集合称为样本（或子样）。比如从某天一机床生产出的另件中取出 50 个，这 50 个另件就是一个样本。从一个总体中可以取出很多种个体集合，因而有很多种样本。特別地，总体是一个特殊的样本。样本中所包含个体的数目称样本容量用 n 表示。那末样本可以用 $x_1, x_2, \cdots, x_n$ 来表示。（注意：$x_1, x_2, \cdots, x_n$ 不了解为总体 $x_1, x_2, \cdots, x_N$ 的前 n 个数）抽取样本的过程称为抽样。从某天一机床生产出的另件中取出一个样本，亦可以說从另件中抽样。

总体，样本都是相对的。同一事物由于研究的問題不同，有时可以作为总体，有时亦可以作为样本。举例来說，某天一車床生产出来的产品全体，我們若檢查此日这車床产品长度的質量，那末这些产品长度就組成一个总体。若檢查整个車間在这天产品长度的質量，那末这些产品长度就是一个样本。

数理統計的目的就是利用抽样的方法得到总体的一个样本，从而依据样本的情况来判断总体的情况。例如說从 1000 个产品中取出 50 个产品进行檢查，从而判断1000个产品中廢品率是多少。从总体中取的样本愈大則样本的情况愈能反映总体的情况，但很大的样本要考查它的性質很困难。若取的样本太小，則不能反映总体的情况。在数理統計中将要解决二个問題：1.如何抽样，抽多少；2.从样本的情况如何来判断总体的情况。

数理統計另一目的是利用抽样方法从几个总体中每个总体得到几个样本，从几个样本的情况判断几个总体的某种性質有否差異，或者判断几个总体之間的联系。

在本节中我們說总体和样本都是一些数的集合，这些数字的分布的情

况如何描繪呢？来下节中将給出描繪数据的方法。

§2. 数据分布——頻数分布与頻率分布

我們假定：在§2，§3的討論中数据 $x_1, x_2, \cdots, x_n$ 可以表示总体的一个样本，亦可以表示总体本身（当 $n=N$ 时，实际上总体是一个特殊样本）。

一般地說实驗观察得到的数据常常是分散的，沒有系統的，我們必須按性質分类，并加以整理，把同类性質的数据归納在一起，并且按大小排列起来。按大小順序排起来的数据称为統計序列，还是用 $x_1, x_2, \cdots, x_n$ 表示，此时 $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$。

表示数据分布情况有两种方法：一种是列表的形式，一种是繪图的形式。从表和图中我們能看出数据的分布情况，并且能粗淺的看出，数据集中在什么地方，集中程度如何。

上节已談到，实驗观察得到的数据按性質分有二种：一种是离散的，亦就是数据从离散分布的总体中取得的。比如只能取非負整数值。另一种是連續的，亦就是数据从連續分布的总体中取得的。比如說产品的长度。

下面分二种情况进行討論：

1. 在数据不多特别是离散的情况下，通常把按大小排列的数中相同的归在一起。設 $x_1, x_2, \cdots, x_n$ 中有 e 个不相同的数，$x_1^*, x_2^*, \cdots, x_e^*$，其中 x_i^* 有 m_i 个 $(i=1, 2, \cdots, e)$。我們称 m_i 为 x_i^* 出现的頻数。所有頻数加起来称为总頻数（即是 n）。頻数 m_i 比上总頻数 n 称为 x_i 出现的頻率。列成表：

	x_1^*	x_2^*	……	x_e^*
頻　数	m_1	m_2	……	m_e
頻　率	$\frac{m_1}{n}$	$\frac{m_2}{n}$	……	$\frac{m_e}{n}$

第一二行称頻数分布，第一三行称頻率分布。

表中 $\sum_{i=1}^{e} m_i = n$，$\sum_{i=1}^{e} \frac{m_i}{n} = 1$。

（讀者可以把数据頻率分布表与概率論中的离散随机变数的分布列作

比較）。

数据的频数分布和频率分布都描繪了数据的分布情况。更明确地說频数分布給出了各个数值个数的分布，频率分布給出了各个数值佔全体数据的百分比的分布。

若用图表示更能直观地显示出数据的分布情况。

作图时，只要把数据作为横坐标频数（频率）作为纵坐标作出点子，連接各点就得到频数(频率)分布图，这种分布图称为多角形频数（频率）分布图。有的书上只画出表示频数（频率）高点的点子，而不作联接各点的直綫。（見图1）

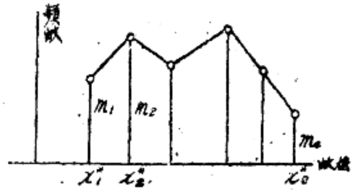

图 1

作多角形频数分布图如右：

若把图中频数 m_i 分别换成频率 $\frac{m_i}{n}$ $(i=1, 2, \cdots, n)$亦即把多角形的頂点的高度压縮 n 倍，则得到多角形频率分布图。

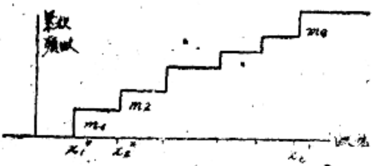

图 2

实用上还有一种所謂累积频数（频率）分布图。对于横坐标上任一值 x 取纵坐标为小于 x 的数据的频数（频率）就得到累积频数（频率）分布图(見图2)。

累积频数（频率）分布图是阶梯形曲綫。在間断点 $x_1^*, x_2^*, \cdots, x_e^*$ 上的跳跃分别为 $m_1, m_2, \cdots, m_e$。

若把图中频数 m_i 分别换成频率 $\frac{m_i}{n}$ $(i=1, 2, \cdots, m_e)$，亦即把图形的高度压縮 n 倍，则得到累积频率分布图。

（讀者可以把累积频率分布与概率論中的离散随机变数的分布函数作比較）。

2. 若数据的个数超过100特别是数据連續的情况下，用上法处理起来是繁复的。可以用分組法来处理，下一节可以看到用分組法处理在計算上可以带来很大方便。所謂分組法即是把数据划分成一些組，划分成一些組后把每一个組作为一个基本单元，在同一組內的数据都看成是相同的，

它們都等于組的兩端的平均数，亦可以說等于組中值，組中值就作为一組中数据的代表。因而組的个数比之于不同的数据的个数要少得多，这样处理起来就方便多了。現把 n 个数据划分成 e 个組（为以后方便起見，亦用 e 表示組的个数）。

$$\Big|\ x_1,\quad x_2,\ \Big|\ \cdots\cdots,\quad x_n\ \Big|$$
$$a_0 \qquad\qquad a_1 \qquad\qquad a_e$$

兩相邻分点間距离 a_1-a_0，a_2-a_1，…，a_e-a_{e-1} 称組长（或組距）。在实际工作中分組的原則如下（这些原則完全由实际工作經驗所得）：

分組的原則：

(1) 数据需要在一百个以上。

(2) 在一般情况下把数据分成 15 組到 25 組为宜，但至少不能少于10組。若組分得太多，就不方便。若組分得太少，就太粗糙，不能很好反映实际情况。

(3) 在一般情况下分組分得每組长都相同。在数据分布极不均匀时，在分組时要求每个組內所包含的数据的个数不宜相差太悬殊。亦就是在划分，組的长度时在数据較密的地方可以取短一些，在数据較疏的地方可以取长一些。

(4) 在兩組的交界上的数据約定归于右边一边組。

同样地每組內的数据的个数称为頻数。所有頻数的和称为总頻数即是 n。各組的頻数比上总頻数称为頻率。列成表：

組	(a_0a_1)	(a_1a_2)	……	(a_{e-1},a_e)
頻　数	m_1	m_2	……	m_e
頻　率	$\frac{m_1}{n}$	$\frac{m_2}{n}$	……	$\frac{m_e}{n}$

第一二行称頻数分布，第一三行称頻率分布。

对于分組情形，可用直方頻数（頻率）分布图描繪分組的頻数（頻率）分布情况。直方图有二种：

(i) 在横坐标上画出点 a_0，a_1，a_2，…，a_e，以每一組为底边，相应組上的頻数（頻率）为高作矩形，得到高直方頻数（頻率）分布图。

作直方頻数分布图如下（图 3）：

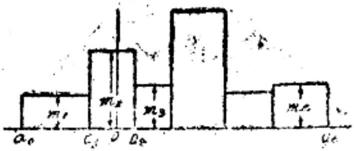

图 3

若把图中频数 m_i 分别换成频率 $\frac{m_i}{n}$ $(i=1, 2, \cdots, n)$ 亦即把各矩形的高度压縮 n 倍，则得到直方频数（频率）分布图。

这种直方图的缺点在于假如一个組分得长一些，那末在这組上的频数（频率）就要大一些，結果势必相应的矩形要高一些。这样不能很好看出频数（频率）分布情况，把各組的频数（频率）加以比較。为了避免这一点，可以作另一种直方图。

(ii) 在横坐标上画出点 a_0、a_1、a_2、$\cdots$，a_e。以每一組就底边，矩形高取为相同应組上的频数（频率）被組的长度除得的数值，这数值等于組内平均分布，在单位长度区間的频数（频率）的大小。这样表示频数（频率）分布情况要可避免掉因組的长度大，而相应組上的频数（频率）亦大这一点。因为各个矩形高 $=\frac{\text{频数（频率）}}{\text{底长}}$，可見矩形面积就等于相应組上的频数（频率）了。

作直方频数分布图如下（見图 4）。图中所有矩形面积之和等于总频数 n。

特别要提出的是直方频率分布图（图 5）：

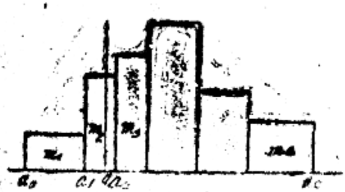

图 4

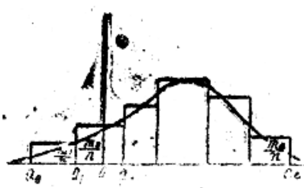

图 5

它只是把直方频数分布图的高压縮 n 倍。图中所有矩形面积的总面积等于 1。在数据連續的情况下，从理論上讲，当 n 不断增大，每一組的长度分得愈来愈小时，表示频率分布的阶梯形曲綫就趋近于一条光滑的曲綫，曲綫記为 $f(x)$。曲綫 $f(x)$ 称为频率分布曲綫。函数 $f(x)$ 称为频率分布密度函数，簡称频率分布密度。

频率分布密度理論上描繪了連續数据的分布情况。

（讀者可以把数据的频率分布密度与概率論中的連續随机变数的分布密度作比較）。

频率分布密度有如下性質：

(1) $f(x)\geq 0$。

(2) 数据分布在区間 $[x, x+dx)$ 中的频率等于区間 $[x, x+dx)$ 之上曲綫 $f(x)$ 之下梯形的面积 $f(x)dx$（見图6）。

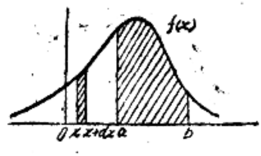

图 6

由(2)可得

(3) 数据分布在区間 $[a, b]$ 中的频率等于区間 $[a, b]$ 之上曲綫 $f(x)$ 之下曲边梯形的面积 $\int_a^b f(x)dx$（見图6）。

(4) 曲綫 $f(x)$ 下的总面积 $\int_{-\infty}^{+\infty} f(x)dx=1$ 等于1。

在分組的情况同样可作出累积频数（频率）分布图。把在組內的频数（频率）看成均勻地分布着，因而在組內 x 增加时累积频数（频率）就均勻地增加，故在組內的图形是直綫，累积频数分布如下（見图7）：

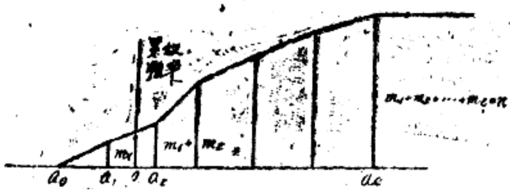

图 7

特别値得提出的是累积频率分布图。图形如下：

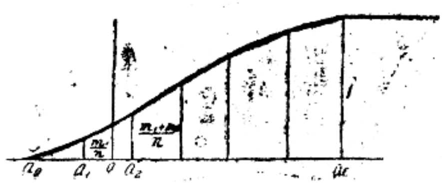

图 8

它是把累积频数分布图高度压縮 n 倍。在数据連續情况，当 n 不断增

大，分的组的長度愈来愈小时，在理論上累积頻率曲綫趋近于一条光滑的曲綫。記为 $F(x)$。曲綫 $F(x)$ 称为累积頻率分布曲綫(見图8)。函数 $F(x)$ 称为累积頻率分布函数。(讀者可与概率論中的連續随机变数的分布函数作比較)。累积頻率分布函数描繪了数据的分布情况。

頻率分布密度和累积频率分布函数都描繪了数据的分布情况，那么它們之間又有怎样的关系呢？在数据連續分布时，有一满意的回答。

现在来說明頻率分布函数 $f(x)$ 和累积頻率分布函数 $F(x)$ 之間的关系：

在数据連續分布时，数据落在区間 $[x, x+dx)$ 內的頻率为 $F(x+dx)-F(x)$，另一方面它又等于 $f(x)dx$。

$$\therefore \qquad F(x+dx)-F(x)=f(x)dx$$

$$\frac{F(x+dx)-F(x)}{dx}=f(x)$$

因而得到
$$F'(x)=f(x).$$

此式說明累积頻率分布函数的导数等于頻率分布密度。那么利用牛頓—萊布尼兹公式，反过来就有

$$F(b)-F(a)=\int_a^b f(x)dx.$$

引进了頻率分布函数与累积頻率分布函数后，我們把处理一批数据問題变成处理已經用几何图形表示出来的一个函数的問題。从此以后我們就可以用代数和微积分的工具了。

"$n\to\infty$"时数据的頻率分布密度或累积頻率分布函数，在实际問題中反映了无限大总体或近似无限大总体的数据的頻率分布。实际工作中，我們总是通过总体的一个样本的数据分布去推判总体的数据分布，当所取的样本愈大，样本的数据分布愈接近总体的数据分布。由于人們受到时間、精力、經济等各方面的限制，我們得到的样本容量不可能太大。因而"$n\to\infty$"只是数学上的假想，我們假想"$n\to\infty$"时的极限情形样本的頻率分布就变成总体的頻率分布了。由此得出結論：頻率分布密度或累积頻率分布函数从理論上客观地反映了无限大或近似无限大总体的数据的頻率分布情况。有的书上把頻率分布密度和累积頻率分布函数称之为数据的理論分布。

数据的頻率分布和概率論中随机变数的概率分布相类似。从意义上講

是不同的，数据的頻率分布是表示数据的分布情况，那里多一些那里少一些，而随机变数的概率分布表示随机变数取值的概率分布情况。但是，它們之間有着紧密的联系，这一点我們在第4节中将較詳細地闡述。

类似于随机变数的数字特征，在下一节中我們要講数据頻率分布的数字特征。

§3. 数据頻率分布的数字特征

数据頻率分布的数字特征是指納頻率分布的特点的数值。大致地来說可以分二种：一种是数据集中在什么位置，即数据圍繞什么位置分布的。另一种是数据对于集中位置的集中程度如何，它們的定义完全类似于随机变数的数学特征。

1．平均数。

把所有数据相加取平均所得到的数值称为平均数，記为 $\bar{x}$，用式子表示成

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}$$

在不分組的情况：

$$\bar{x}=\frac{m_1x_1^*+m_2x_2^*+\cdots+m_sx_s^*}{n}$$

$$=\frac{m_1}{n}x_1^*+\frac{m_2}{n}x_2^*+\cdots+\frac{m_s}{n}x_s^*$$

在分組的情况：設第 i 組組中值为 x_i^*，設 x_i^* 为第 i 組的代表（$i=1, 2, \cdots, s$），則

$$\bar{x}\approx\frac{m_1x_1^*+m_2x_2^*+\cdots+m_sx_s^*}{n}$$

$$\approx\frac{m_1}{n}x_1^*+\frac{m_2}{n}x_2^*+\cdots+\frac{m_s}{n}x_s^*$$

当然这样算得的平均数与真正的平均数是有誤差的，但是当組分得不太大时，这誤差是相当小的。

平均数是数据分布的很重要的一个特征。它把数据之間的差异平均掉了。数据以平均数为集中中心，圍繞着它分布。在产品检查和测量理論中是一个重要指标。

例題：

x_i^*	m_i	$m_i x_i^*$
19	1	19
20	3	60
21	5	105
22	7	154
23	6	138
24	2	48
25	1	25

$$\sum m_i = 25 \qquad \sum m_i x_i^* = 549$$

$$\bar{x} = \frac{549}{25} = 21.96$$

在組长相等分組的情况，有时用下面的公式計算有平均数是相当麻烦的。下面介紹一个計算平均数的簡单方法。

平均数簡算法：先找出一个在数据分布中佔比較中間的位置的中值 x_0，每个組中值 x_i^* 減去 x_0，再除以組的长度 c，則得到

$$u_i^* = \frac{x_i^* - x_0}{c}.$$

对 u_i^* 取平均得到 $\bar{u}$，$\bar{u}$ 与 $\bar{x}$ 的关系如下：

$$\bar{x} = \sum_{i=1}^{s} \frac{m_s}{n} x_i^* = \sum_{i=1}^{s} \frac{m_s}{n} (cu_i^* + x_0)$$

$$= c\sum_{i=1}^{s} \frac{m_s}{n} u_i^* + x_0 \sum_{i=1}^{s} \frac{m_s}{n} = c\bar{u} + x_0$$

故由 $\bar{u}$ 很簡单地能算出 $\bar{x}$。

例題：

組	x_i^*	m_i	u_i^*	mu_i^*
189.5～193.5	191.5	2	−7	−14
193.5～197.5	195.5	4	−6	−24
197.5～201.5	199.5	7	−5	−35
201.5～205.5	203.5	12	−4	−48
205.5～209.5	207.5	19	−3	−57
209.5～213.5	211.5	24	−2	−48
213.5～217.5	215.5	27	−1	−27
217.5～221.5	219.5	35	0	0

221.5～225.5	223.5	26	1	26
225.5～229.5	227.5	21	2	42
229.5～233.5	231.5	18	3	54
233.5～237.5	235.5	13	4	42
237.5～241.5	239.5	6	5	30
241.5～245.5	243.5	5	6	30
245.5～249.5	247.5	2	7	14
249.5～253.5	251.5	1	8	8
		$\sum m_i=222$		$\sum m_i u_i^*=3$

表中取中間值 219.5 为 x_0，c 的值为 4。

$$\bar{u}=\frac{3}{222}=0.0135.$$

$$\bar{x}=4\times0.0135+219.5=219.554.$$

在数据連續分布时，当数据无限增多，分組分得愈来愈小时，平均数可表成积分

$$\bar{x}=\int_{-\infty}^{+\infty}xf(x)dx$$

其中 $f(x)$ 为数据的频率分布。

2. 中位数。

把数据按大小次序排列起来，最中間的一个数称为中位数。記为 M_e。

数据有奇数个，那末中位数取中間的一个数。例：数据60、80、90、95、98，$M_e=90$。

数据有偶数个，那末中位数取中間二个数的平均值。例：数据 60、80、90、95、98、100、$M_e=\frac{90+95}{2}=92.5$。

中位数亦描写数据的集中位置，数据围绕着中位数分布着。中位数計算很方便，比之于平均数的計算簡单得多。它在产品質量控制中要用到。

3. 均方差。

数据与平均数之差的平方的平均值开根称为均方差，記为 σ。用式表示为

$$\sigma=\sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}}$$

在不分組的情况：

$$\sigma = \sqrt{\frac{m_1(x_1^* - \bar{x})^2 + m_2(x_2^* - \bar{x})^2 + \cdots + m_e(x_e^* - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{m_1}{n}(x_1^* - \bar{x})^2 + \frac{m_2}{n}(x_2^* - \bar{x})^2 + \cdots + \frac{m_e}{n}(x_e^* - \bar{x})^2}$$

在分組的情况：第 i 組内的数据以組的中間值 x_i^* 为代表($i=1, 2, \cdots, e$)，則

$$\sigma = \sqrt{\frac{m_1(x_1^* - \bar{x})^2 + m_2(x_2^* - \bar{x})^2 + \cdots + m_e(x_e^* - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{m_1}{n}(x_1^* - \bar{x})^2 + \frac{m_2}{n}(x_2^* - \bar{x})^2 + \cdots + \frac{m_e}{n}(x_e^* - \bar{x})^2}$$

若均方差愈小，則数据对于平均数之差的平方的平均值愈小。因而数据对平均数之差的絕对数值都很小，說明数据对于平均数的差異都很小，数据对于平均数 $\bar{x}$ 愈集中。由此可見，**均方差是描繪数据对于平均数差異程度或者集中性程度的数。**

均方差在产品檢查与測量理論中是一个很重要的量。

在計算均方差时常采用公式：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2} = \sqrt{\sum_{i=1}^{e} \frac{m_i}{n} x_i^{*2} - \bar{x}^2}$$

公式推导：

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

$$= \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - 2\bar{x}\frac{\sum_{i=1}^{n} x_i}{n} + \bar{x}^2} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2}$$

$$= \sqrt{\frac{\sum_{i=1}^{e} m_i x_i^{*2}}{n} - \bar{x}^2}$$

利用这公式計算可減少一道減 $\bar{x}$ 的手續，所以計算时有方便之处。

例題：