应用数理统计方法

苏州医学院

一九七六年四月

应用数理统计方法

引　言

对在同一生产条件下生产出来的某一批产品的某种性质进行测量时，发现该批产品的测量结果相互间总不免有所差异；即使对同一件产品在同一条件下重复测量，所得结果也不是完全一样的，存在所谓的测量统计误差。类似上述在实现一系列共同条件下所得观测结果互有差异的现象是属偶然性质的、这类现象在现实世界中几乎普遍存在着。在辐射防护专业工作中同样经常面对这一类现象。在一个测量放射性强度的实验里，即使所有测量条件完全保持相同，在这种情况下对同一放射源进行多次重复测量的计数结果仍是有统计涨落的；现场工作中、在同一条件下取得的一些样品，即使样品分析方法和测量条件保持一样、测得各样品的结果间是互有差异的；用实验动物研究电离辐射的生物效应时、即使接受的照射量相同，在其他条件方面也基本相似的实验动物中所观测到的各别个体的效应很不一致。像这类的例子不胜枚举。概率论与数理统计是一门研究偶然现象的规律性的学科、在研究上述这些具有偶然性质的现象时，须应用建立在该门学科理论基础上的数理统计方法。

统计方法大体上涉及二类问题：统计叙述和统计推断。前者是对实际观测得到的资料进行整理和归纳、用少数几个数字特征（有时并辅以图表）将收集所得资料的主要特征表现出来，便于理解、掌握并作进一步的分析和阐述。后者则属由部分推断全体的问题。

在数理统计中、将研究对象的全体（满足指定条件的所有个体或单位的集合）叫做总体，而将取自总体的一部分实际观测的个体或单位的集合称作样本。例如，在某一时间要测量指定的某一地区土壤的放射性水平、则该地区应测的全部土壤构成总体，而取自该地区的一些土壤样品则为该总体的一个样本。又如假定在同一条件下对一个样品的测量能无限多次地进行下

去，这无限多次测量的结果即为总体；实际上只测量了数次，这数次测量的结果组成了该总体的一个样本。在实际工作中一般只能对样本中各别个体或单位的某种性质进行观测，然而观测的结果却总是要推广到属于该研究对象的总体去，这是一个非常重要的科学推断过程。其实，数理统计的中心问题在于如何根据样本探求有关总体的种种知识，以及从样本取得的资料去检验关于总体的种种假设。

为使能从样本观测结果对总体进行科学的统计推断，重要的条件是要保证样本对总体具有充分的代表性，要避免样本具有偏性。理想的抽样方法是在抽样时使得总体内的每个个体或单位均有同等被抽取的机会，这样组成的样本称为随机样本。统计上常采用一些具体措施（如使用随机数字表，或在简单的情况下用抽签，掷币的方法）以保证抽样的随机性。在数理统计中随机样本是基本和重要的概念，进行统计推断是基于这一概念的基础上的。因此，在环境监测工作中，应根据不同的具体情况，顺应地考虑组成随机样本的方法。

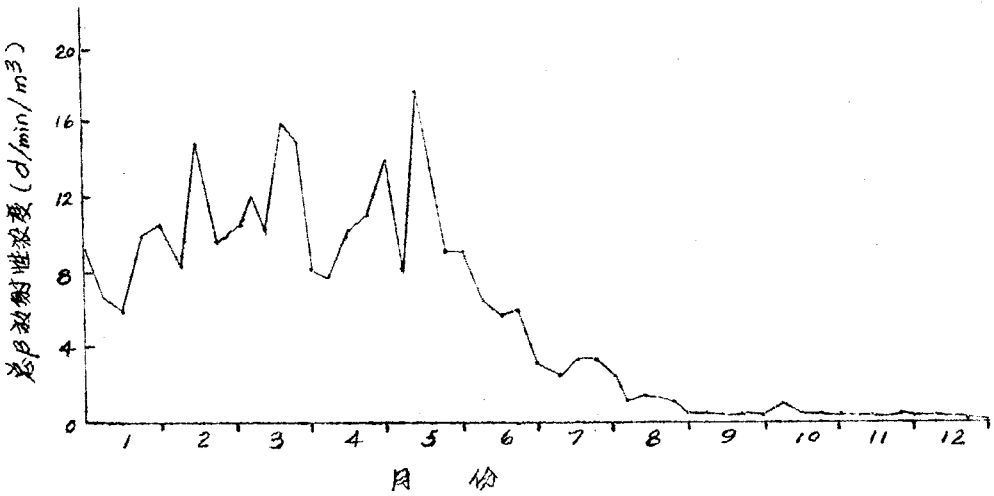
编写的这份材料的内容虽然主要是涉及应用于收集所得的数据的统计分析方法，但同时又该强调指出统计方法在研究设计中的重要性。研究工作中若在制订计划时就有周密的统计上的考虑，则往往能从收集所得的数据中获得更多的信息，达到研究的预期目的，从而使研究工作更有成效。因此，我们应把统计方法的应用不仅看作为处理和分析资料的一种手段，同时也是研究设计的一个有机组成部分。

编写这份材料的目的，在于简要地介绍一般常用的数理统计方法，叙述时尽量指出它在辐射防护工作中的应用，以期引起对数理统计方法的重视和兴趣，便于在今后的工作中充分利用这一有效的手段，将辐射防护工作做得更好。

一，变量分布及其数字特征

通过实际测量收集所得的是一批数值互有差异的数据。分析数据的第一步，是要将收集所得的资料进行整理和归纳、并将结果用表和各的形式列示。

当观测数据中各个数值的大小与出现的一定次序（如按时间上的先后，距离上的远近）有关联时，可直接按出现的次序将各数值列示或在各上点示，以了解观测的指标（下面将称为变量）的数值在时间上的趋势或在不同地点上的特点，如某地点一年内不同时间上放射性水平的变化或某一时间不同距离的地点放射性水平的改变等。各1—1表明某年某采样点空气中总β放射性在时间上的变化趋势（各上数值以每日测定的结果按一周平均的平均数分别点示，单位为 $d/min/m^3$），从各中可见该地该年前五个月空气中放射性要比较高的，而自八月底以后则均低于 1 $d/min/m^3$。



各1-1 某年某采样点空气中总β放射性浓度在时间上的变化趋势

为了解一批变量值总的分布情况，通常要将各变量值按其数值大小重新排列起来。当变量值为数众多时，将它们分成若干组，列示整理和归纳的结果。

1. 频数分布和频率分布

表1—1和表1—2是数据经整理后的二个频数分布的例子。表中 x_i（$i=1,2,\cdots,n$）表示变量出现的数值（已分组数据用组中值表示），f_i 表示 x_i 在全部数据中出现的次数，或称频数。f_i 被总频数 $\sum f_i$ 除所得之值 $f_i/\sum f_i$，即为各变量值（或某一组变量值）在全部数据中出现的频率。

表1—1　放射性物质衰变计数结果（计数/7.5秒）

[illegible] x	频数 f	频率 $f/\sum f$	累积频数 c.f.	累积频率 c.f./$\sum f$
0	57	0.022	57	0.022
1	203	0.078	260	0.100
2	383	0.147	643	0.247
3	525	0.201	1168	0.448
4	532	0.204	1700	0.652
5	408	0.156	2108	0.808
6	273	0.105	2381	0.913
7	139	0.053	2520	0.966
8	45	0.017	2565	0.984
9	27	0.010	2592	0.994
10	16	0.006	2608	1.000
总计	$\sum f=2{,}608$	0.959	—	—

表1—2　某地某年某月空气中 SO_2 浓度分布

日平均浓度 ($\mu g/m^3$)	样品数 f	频率 $f/\sum f$	累积频数 C.f.	累积频率 C.f./$\sum f$
0—	2	0.004	2	0.004
40—	27	0.058	29	0.063
80—	63	0.136	92	0.198
120—	98	0.211	190	0.409
160—	87	0.188	277	0.597
200—	69	0.149	346	0.746
240—	45	0.097	391	0.843
280—	33	0.071	424	0.914
320—	12	0.026	436	0.940
360—	9	0.019	445	0.960
400—	8	0.017	453	0.976
440—	3	0.006	456	0.983
480—	5	0.011	461	0.994
520—	2	0.004	463	0.998
560—600	1	0.002	464	1.000
总计	$\sum f=464$	0.999	—	—

显然，所有频率之和恒等于1。

有时要了解一批数据中小于（或大于）某一数值的变量值出现的频数或频率，这时可将频数或频率分布表中频数或频率栏内的各数值由上往下（或由下往上）相加，求得累积频数或累积频率。

变量分布一般有离散型和连续型之分。在离散型分布中变量只能取数轴上可数个孤立的数值，如表1—1中每单位时间（7.5秒）的放射性计数为0，1，2，…等6；在连续型分布中变量可以取数轴的某个区间上的一切数值，如表1—2中各次测定结果的单位是以$\mu g/m^3$表示的，这并不意味着测定结果确切地为某一数值，该数决定于测定方法的准确度。

在频数或频率分布表上已能较明显地反映出变量分布的情况，如那些变量值出现的频数较多或频率较大。若将表内资料绘制成图，则变量分布的情况更为直观醒目。对离散型分布的资料通常在图中用条的高度表示频数的多寡（图1—2），而对连续型分布的资料则用长方形面积的大小表示频数的多少（图1—3）；前一种图形称为条图，后一种图形称为直方图。若将频数分布图中纵轴的尺度除以总频数，则得变量的频率分布图。在频率分布图上，条图中诸长条高度的总和等于1，各长条高度分别表示在总高度中所占的百分比；直方图中诸长方形面积的总和等于1，各长方形的面积分别表示在总面积中所占的百分比。

2．样本分布和总体分布

若抽样时遵循随机的原则，由实测数据得到的频率分布所表明的是一个随机样本的变量分布情况，而样本所取自的总体中的变量分布情况是未知的。显然，样本分布不会与总体分布完全相同，分析样本变量分布的目的是要推测未知总体的变量分布情况。样本分布一般也称为经验分布，而将相应的总体分布称为理论分布。

引用概率论中的大数定律可知，若样本中所含变量值的个数众多时，由样本的经验频率就给出总体相应的理论频率的近似值。以后，我们将此理论频率称为概率，并以符号P表示。在这份模拟教学讲义的材料内将不对概率作进一步讨论。

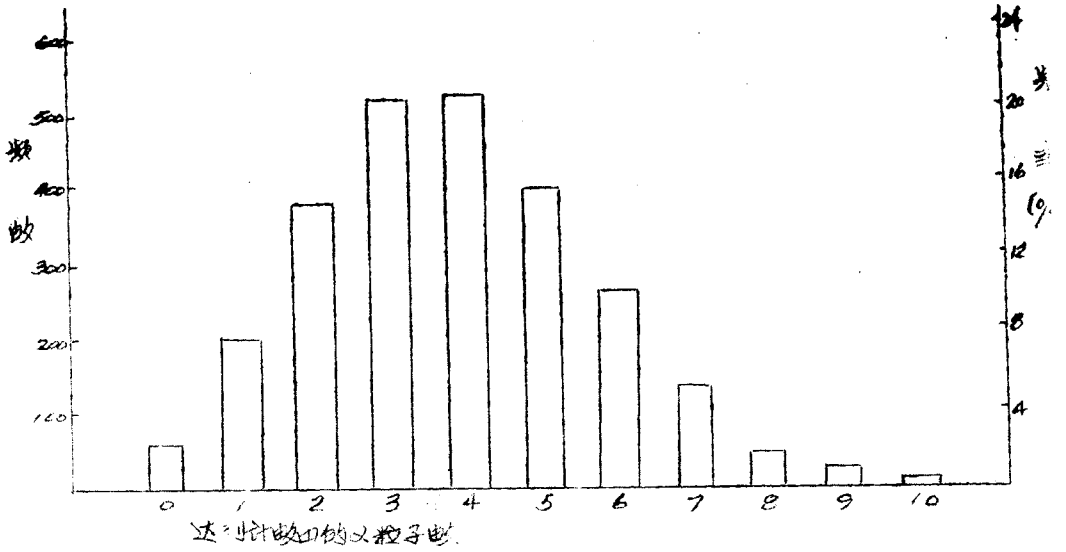


图1-2 放射性物质发射X粒子数的频数(频率)分布

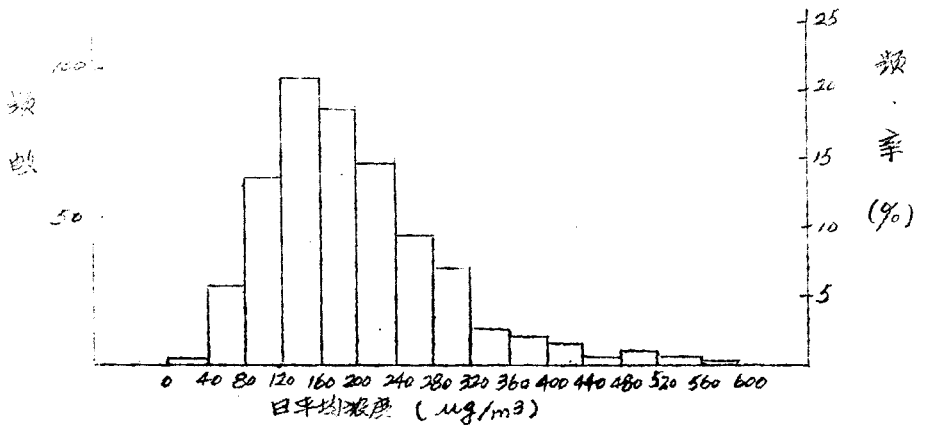


图1-3 某地某年某月空气中SO₂浓度频数(频率)分布

在离散型分布中，设随机变量ξ所取可能的取值为x_1、x_2、…，则样本中变量值x_i（$i=1, 2, \cdots$）出现的频率为$f_i/\Sigma f$，在总体中随机变量ξ取值x_i的概率可写作

$$P\{\xi = x_i\} = P_i$$

显然，$\sum_i P_i = 1$（$i = 1, 2, \cdots$）。现将

$$F(x)=P\{\xi<x\}=\sum_{x_i<x}P\{\xi=x_i\}=\sum_{x_i<x}P_i$$

称为随机变量 ξ 的分布函数（相当于样本中的累积频率），知果对于所有 x 值知道 $F(x)$，也就知道了 ξ 的概率分布。如

$$P\{x_1\leq\xi<x_2\}=P\{\xi<x_2\}-P\{\xi<x_1\}$$
$$=F(x_2)-F(x_1)$$

在连续型分布中，只能说随机变量 ξ 所取的值在数轴的某个区间上，于是样本中随机变量的数值表现在该区间上的频率为 $f_i/\Sigma f$，在直方图中该频率即为该区间上长方形的面积占所有长方形面积总和的百分比。现设不断增大样本容量，即不断增加观测的总频数，同时不断缩小数据分组的组距，这时直方图中小长方形的数目也就不断增加。当样品容量无限增大且组距无限缩小时，小长方形的顶边逐渐趋近于一条光滑的曲线（图1-4），这条曲线就是理论分布的曲线。因此，总体中随机变量落在某一区间上的概率，在图形上可用理论分布曲线下面介于该区间的横坐标之间的面积占曲线下总面积的百分比来表示。

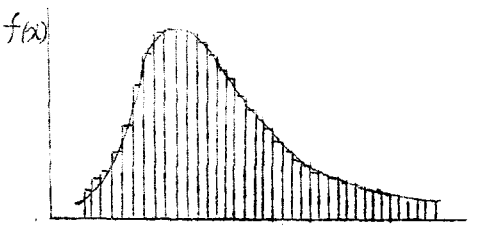


图1-4　理论分布示意

若以 $f(x)$ 表示理论分布曲线的函数式，则介于横轴上 x_1 与 x_2 之间（$x_1<x_2$）的面积 A 可用下式表示：

$$A=\int_{x_1}^{x_2}f(x)dx$$

现令曲线下的总面积等于 1，则 A 表示随机变量落在区间 (x_1, x_2) 上的概率，即

$$P\{x_1<\xi<x_2\}=\int_{x_1}^{x_2}f(x)dx$$

式中 $f(x)$ 表示随机变量 ξ 在点 x 处的概率密度，连续型随机变量的分布规律一般由它的概率密度 $f(x)$ 给出。显然

$$P\{-\infty<\xi<+\infty\}=\int_{-\infty}^{\infty}f(x)dx=1$$ （在数理统计中常将积分上下限写为$-\infty$和$+\infty$），随机变量ξ的分布函数$F(x)$（相当于样本中的累积频率）则为

$$F(x)=P\{\xi<x\}=\int_{-\infty}^{x}f(x)dx,$$

由$F(x)$可以得出ξ落在任意区间(x_1,x_2)上的概率为

$$P\{x_1<\xi<x_2\}=F(x_2)-F(x_1)=\int_{-\infty}^{x_2}f(x)dx-\int_{-\infty}^{x_1}f(x)dx$$
$$=\int_{x_1}^{x_2}f(x)dx.$$

不难知道，概率密度$f(x)$就是ξ的分布函数$F(x)$的导数，即$F'(x)=f(x)$

3，分布的数字特征.

通过实际观测得到的是一批数值参差不齐的数据，为了解这批数据所具有的某些重要特征，需要用一些数字来反映这些特征，这种数字称为特征数。常用的数字特征有二类，一类是表示分布的集中位置的（如平均数、中位数、众数等），一类是表示分布的离散程度的（如极差、标准差等）。

在实际工作中，通常是从分析样本数据入手的。在保证随机取样的条件下，一批实测数据即为一个随机样本。统计上把由样本数据得到的数字特征称为统计量，而将总体的相应的一些数字特征称为参数。显然，总体参数是一些常数，而样本统计量随抽取样本的不同是一些变量。下面先介绍一些常用的样本数字特征。

1）平均数　又称算术平均数，是表明分布集中位置最常用的一个数字特征。以$\bar{x}$表示样本平均数，其计算公式可写成

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n}$$，（未分组数据）（式1—1）

式中n表示变量值的个数，又称样本容量.

$$\bar{x}=\frac{f_1x_1+f_2x_2+\cdots+f_nx_n}{f_1+f_2+\cdots+f_n}=\frac{\sum_{i=1}^{n}f_ix_i}{\sum_{i=1}^{n}f_i}$$.（已分组数据）（式1—2）

式中f_i表示变量x_i的频数，$\sum_{i=1}^{n}f_i$为总频数。

如由表1—1资料用式1—2可求得

$$\bar{x} = \frac{57\times 0 + 203\times 1 + \cdots\cdots + 16\times 10}{2608} = \frac{10086}{2608} = 3.9\ 175\text{秒}$$

对表1—2的资料，可用各组的组中值代表归入各组的诸变量值的数值，然后用上式进行计算。为运算简便起见，需先将诸变量值x_i减去一个常数x_0并除以一个常数c，所得结果用u_i表示，即令

表1—3　　用表1—2资料计算均数和标准差

日平均浓度分组 (mg/m³)	组中值 x	频数 f	u	fu	fu^2
0 —	20	2	−6	−12	72
40 —	60	27	−5	−135	675
80 —	100	63	−4	−252	1008
120 —	140	98	−3	−294	882
160 —	180	87	−2	−174	348
200 —	220	69	−1	−69	69
240 —	260	45	0	0	0
280 —	300	33	1	33	33
320 —	340	12	2	24	48
360 —	380	9	3	27	81
400 —	420	8	4	32	128
440 —	460	3	5	15	75
480 —	500	5	6	30	180
520 —	540	2	7	14	98
560 — 600	580	1	8	8	64
总　计	—	464	—	−753	3761

$$u_i = \frac{x_i - x_0}{c}$$

求得变量u的均数$\bar{u}$后，再用下式得到$\bar{x}$：

$$\bar{x} = x_0 + c\bar{u} \qquad (式1-3)$$

式中x_0为任一组的组中值（通常将频数f_i最大一组的组中值定为x_0），c为等距分组的组距。

现用式1—3求得表1—2资料的均数为

$$\bar{x} = x_0 + c\bar{u} = x_0 + c\cdot\frac{\Sigma fu}{\Sigma f} = 260 + 40\times\frac{(-753)}{464} = 195.1\ mg/m^3$$

2）中位数　将一批变量值按数值由小到大排列以后，位于中间位置上的变量值即为中位数。中位数以Me表示。如果变量值的个数为偶数，取中间位置上相邻二个变量值的平均数为中位数。

对于已分组的连续变量资料，可用内插法按下式计算中位数：

$$M_e = L_{Me} + \left(\frac{n}{2} - n_1\right) \cdot \frac{C}{f_{Me}} \qquad (式1-4)$$

式中 L_{Me} 为中位数所在组的下限，n 为样本的总量，n_1 为中位数所在组以前各组频数的和，C 为中位数所在组的组距，f_M 为该组的频数。

如表1-2资料的中位数为：

$$M_e = 160 + \left(\frac{164+1}{2} - 119\right)\frac{40}{87} = 160 + 19.54 = 179.5\ mg/m$$

3）众数　样本中出现次数最多的变量值即为众数。众数用 M_o 表示。对于已分组的连续变量资料，可用下式求 M_o 的近似值。

$$M_o = L_{Mo} + \left(\frac{d_1}{d_1 + d_2}\right) \cdot C \qquad (式1-5)$$

式中 L_{Mo} 为众数所在组的下限，d_1 为众数所在组频数减去前一个组频数的差数，d_2 为众数所在组频数减去后一组频数的差数，C 为众数所在组的组距。

如表1-2资料的众数为：

$$M_o = 120 + \frac{93-63}{(93-63)+(93-87)} \cdot 40$$

$$= 120 + 33.33$$

$$= 153.4\ mg/m^3$$

4）几何平均数　其定义为

$$G = \left\{\prod_{i=1}^{n} x_i\right\}^{\frac{1}{n}} \qquad (式1-6)$$

式中 G 为几何平均数，x_i 为变量值，n 为变量值的个数。上式表示 n 个变量值的几何平均数等于它们的连乘积的 n 次方根。实际上是用下式进行计算的：

$$\log G = \frac{1}{n}\sum_{i=1}^{n} \log x_i \qquad (式1-7)$$

据报导，环境介质中许多物质浓度的分布近似呈以后将提到的对数正态分布，这时计算和应用几何平均数有重要意义。对于大数量资料，可将原变量值先变换成对数并进行分组以后，按式1－3类似的方法计算 $\log G$

5）标准差和方差

前面提及的一些数字特征是表示分布的集中位置的，在表示分布的离散程度的数字特征中最常用的是标准差和方差。

样本方差用 S^2 表示，一般采用下式计算：

$$S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \qquad (式1-8)$$

实际运算时分式中的分子（称离均差平方和）常按下式进行计算：

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - \frac{(\sum_{i=1}^{n}x_i)^2}{n} \qquad (式1-9)$$

计算样本方差时所以用 $n-1$ 作分母而不用 n，其原因将在后面提到。

对已分组资料，样本方差可按下式计算：

$$S^2 = \frac{\sum_{i=1}^{n}f_i x_i^2 - \frac{(\sum_{i=1}^{n}f_i x_i)^2}{\sum_{i=1}^{n}f_i}}{\sum_{i=1}^{n}f_i - 1} \qquad (式1-10)$$

方差的平方根称标准差，用 S 表示，标准差在统计分析中应用极广。标准差是有量度单位的数字特征，其单位与变量的单位相同。标准差恒取正值，不取负值，其数值大小表明分布的离散程度大小。

计算标准差时，若先将诸变量值 x_i 减去一个常数 x_0，所得标准差的数值不变！即将样本分布在数轴上向左平移 x_0 个单位，其离散程度不变）；若先将诸变量值 x_i 除以一个常数 c，则所得的标准差数值为原标准差的 $\frac{1}{c}$。利用这一特点，可使计算简便。如对已等距分组的资料（如表1－2），令 $u = \frac{x_i - x_0}{c}$，先求得新变量 u 的标准差 S'，再乘上 c 即得所求的标准差 S。计算式子如下：

$$S = S' \times C = \sqrt{\frac{\sum_{i=1}^{n} f_i u_i^2 - \frac{\left(\sum_{i=1}^{n} f_i u_i\right)^2}{\sum_{i=1}^{n} f_i}}{\sum_{i=1}^{n} f_i - 1}} \times C \qquad (式1-11)$$

式中 C 为等距分组的组距。当总频数 Σf_i 很大时，计标公式可写成：

$$S = \sqrt{\frac{\sum_{i=1}^{n} f_i u_i^2}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i}\right)^2} \times C \qquad (式1-12)$$

如由表1-2资料，可求得：

$$S = \sqrt{\frac{3761}{464} - \left(\frac{753}{464}\right)^2} \times 40$$

$$= 93.6$$

有时还偶尔遇到均差这一数字特征，均差用 M.D. 表示，计标公式如下：

$$M.D. = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \qquad (式1-13)$$

由上式可知，均差也是表示分布函数程度的数字特征，但它在进一步统计分析中的应用不如标准差。

6）极差　观测数据中最大值与最小值之差，用 R 表示，即：

$$R = x_{max} - x_{min} \qquad (式1-14)$$

式中 x_{max} 为样本数据中的最大值，x_{min} 为最小值。

极差的计标极为简单，在一定条件下，用样本极差 R 代替样本标准差 S 作为表示离散程度的数字特征，在使用时甚为方便（参阅第三章控制图的应用）。

7）变异系数　为衡量分布的相对离散程度，常用标准差与平均数的百分比值这一无量纲的数字来表示，这一比值称为变异系数，用 C.V. 表示，即

$$C.V. = \frac{S}{\bar{x}} \times 100\ (\%) \qquad (式1-15)$$

除上述表示分布集中位置和离散程度的二类数字特征外，还有表示分布不对称程度的偏度数字特征和表示分布为高狭峰或平坦峰的峰度数字特征，本材料内对后二类数字特征将不作进一步阐述。

8）数学期望

前面曾以表1—1资料用 $\bar{x}=\frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$（式1-2）求得变量样本分布的平均数，

这一平均数也就是变量值 x_i 分别以其频数 f_i 为权得到的加权平均数。式1—2可改写成：

$$\bar{x}=\sum_{i=1}^{n}\left(x_i \cdot \frac{f_i}{\sum_{i=1}^{n} f_i}\right),$$

即加权平均数等于变量值 x_i 与其频率 $\frac{f_i}{\sum f_i}$ 乘积的总和。

现将加权平均数的概念应用于总体分布，就得到数学期望的概念。

设离散型随机变量 ξ 的概率分布为

$$P\{\xi=x_i\}=p_i, \qquad i=1,2,\cdots,n$$

则加权平均数

$$\frac{x_1p_1+x_2p_2+\cdots+x_np_n}{p_1+p_2+\cdots+p_n}=\frac{\sum_{i=1}^{n} x_ip_i}{\sum_{i=1}^{n} p_i}=\sum_{i=1}^{n} x_ip_i$$

称为随机变量 ξ 的数学期望，并记为 $E[\xi]$。显然，$E[\xi]$ 就是该总体分布的平均数。现用 μ 表示总体平均数，则

$$\mu=E[\xi]$$

同理，设连续型随机变量 ξ 的概率密度为 $f(x)$，则随机变量 ξ 的数学期望为

$$\mu=E[\xi]=\frac{\int_{-\infty}^{\infty} xf(x)dx}{\int_{-\infty}^{\infty} f(x)dx}=\int_{-\infty}^{\infty} x\,f(x)\,dx$$

可见，一个随机变量的数学期望就是这个随机变量的所有可能值以其相应的概率或概率密度为权的加权平均数。

同理，现用 σ^2 表示总体方差，在离散型分布中有

$$\sigma^2 = E\left[(\xi-\mu)^2\right] = \sum_{i=1}^{n}\left\{x_i - E[\xi]\right\}^2 \cdot P_i$$

在连续型分布中有

$$\sigma^2 = \int_{-\infty}^{\infty}\left\{x - E[\xi]\right\}^2 f(x)\,dx$$

在数理统计中，将 $E[\xi^K]$ 称作随机变量 ξ 的分布关于原点的第 K 阶矩（$K=0,1,2,\cdots$），简称 K 阶原点矩；将 $E\{x-E[\xi]\}^K$ 称作关于平均数的第 K 阶矩，简称 K 阶中心矩。前面已证，一阶原点矩就是总体分布的平均数 μ，二阶中心矩为总体分布的方差 σ^2。中心矩与原点矩之间存在有一定的关系如不难证明：

$$\sigma^2 = E\left\{x-E[\xi]\right\}^2 = E[\xi^2] - \left\{E[\xi]\right\}^2$$

顺便指出，随机样本的数字特征（$\bar{x}$，S^2 等）也是随机变量，同样可以求得 $\bar{x}$ 和 S^2 的数学期望。通过运算可以证明：

$$E[\bar{x}] = E[\xi] = \mu.$$

即样本平均数的数学期望等于原随机变量的数学期望，也就是等于总体平均数。按数理统计学中的定义：若估计值 $\hat{\theta}$ 的数学期望等于被估计值 θ，则称 $\hat{\theta}$ 为 θ 的无偏估计；因此，样本平均数 $\bar{x}$ 为总体平均数 μ 的无偏估计。通过运算也可以证明：

$$E[S^2] = E\left[\frac{\sum(x-\bar{x})^2}{n-1}\right] = \sigma^2$$

即样本方差 $S^2 = \frac{\sum(x-\bar{x})^2}{n-1}$ 的数学期望等于原总体分布的方差 σ^2，也就是说，$S^2 = \frac{\sum(x-\bar{x})^2}{n-1}$ 是 σ^2 的无偏估计。若用 $S^2 = \frac{\sum(x-\bar{x})^2}{n}$ 作为 σ^2 的估计值，可以证明：

$$E\left[\frac{\sum(x-\bar{x})^2}{n}\right] = \frac{n-1}{n}\sigma^2 < \sigma^2$$

因此，用 $\frac{\sum(x-\bar{x})^2}{n}$ 作为总体方差 σ^2 的估计值，平均来说，较失之过小；所以，一般就用式 1-8 计算样本方差 S^2。

二，几种理论分布

对于一种偶然现象、通过大量观察或测量以后，会发现"大量现象"的规律、这种规律可用经验频率分布来表述。概率论和数理统计学中有一些理论分布用于阐明一些随机变量的概率分布，只要所研究的偶然现象符合应用这些理论分布的要求，将这些理论分布应用于所研究的偶然现象显然是有重要意义的。

1，二项分布

二项分布是一种离散型随机变量的理论分布。

设已知某种结果出现的概率为 p，不出现的概率为 $q=1-p$，若一次结果出现与否和另一次结果出现与否是相互独立的，则在 n 次观察中出现该种结果次数 $\xi=r$（$r=0、1、2、\cdots n$，）的概率为

$$P\{\xi=r\}=C_n^r p^r q^{n-r} \qquad (式2-1)$$

因为 $P\{\xi=r\}$ 刚好是二项式 $(q+p)^n$ 的展开式中的有关项，因此把 $P\{\xi=r\}$ 随 r 的这一概率分布叫做二项式分布。式中 C_n^r 为组合的符号，$C_n^r=\frac{n!}{r!(n-r)!}$

若要知道 n 次观察中出现某种结果的次数小于 x 次的概率，可由其分布函数求得：

$$F(x)=P\{\xi<x\}=\sum_{0\le r<x} C_n^r p^r q^{n-r} \qquad (式2-2)$$

二项分布的随机变量 ξ 的平均数 μ 和方差 σ^2 分别为

$$E\{\xi\}=np \qquad (式2-3)$$

$$\sigma^2=E\{\xi^2\}-\{E\{\xi\}\}^2=npq \qquad (式2-4)$$

于是，随机变量 ξ 的标准差为

$$\sigma=\sqrt{npq} \qquad (式2-5)$$

在一些数理统计用表中，列出在给定 p 和 n 的数值时二项分布中的 $P\{\xi=r\}$ 和 $F(x)$ 值。

2. 泊松分布

二项分布中当 $n\to\infty$，$p\to 0$ 而 np 为有限数时，可得出下述概率分布

$$P\{\xi=r\}=e^{-\mu}\cdot\frac{\mu^r}{r!} \qquad (式2-6)$$

其中μ为分布的参数，e为自然对数的底。这一概率分布叫做泊松分布。其分布函数为

$$F(x)=P\{\xi<x\}=\sum_{0\le r<x}e^{-\mu}\cdot\frac{\mu^r}{r!} \qquad (式2-7)$$

泊松分布的随机变量ξ的平均数和方差分别为

$$E\{\xi\}=\mu \qquad (式2-8)$$

$$\sigma^2=E\{\xi^2\}-\{E\{\xi\}\}^2=\mu \qquad (式2-9)$$

于是，随机变量ξ的标准差为

$$\sigma=\sqrt{\mu} \qquad (式2-10)$$

可见，泊松分布的方差等于其平均数，或标准差等于其平均数的平方根，是泊松分布的重要特点。

像放射性衰变这类偶然现象有以下特征：①所有原子中每个原子在任一指定的时间间隔内衰变的概率是相同的；②所有的原子都是相互独立的，即一个原子在某一时间间隔内发生衰变并不影响其他原子在同一时间间隔内衰变的概率；③由于平均寿命远大于实际测量时间，每个原子在所有相等长度的时间间隔内的衰变的概率是相同的，所以放射性衰变这类偶然现象是服从一项分布这一统计规律的。设在$t=0$时，放射性原子的总数为n，在$t=t$时间内衰变掉一部分，因此可以把n个原子分成衰变掉和没衰变掉的两组。利用式2—1可以得到，在时间从0到t内，衰变掉r个粒子的概率为

$$P(r)=\frac{n!}{r!(n-r)!}(1-e^{-\lambda t})^r(e^{-\lambda t})^{n-r} \qquad (式2-11)$$

式中λ为放射性元素的衰变常数，假如满足$\lambda t\ll 1$（即测量时间$t\ll$半衰期），$n\gg 1$，$r\ll n$三个条件。也就是满足$p\to 0$（$p=1-e^{-\lambda t}\simeq 1-(1-\lambda t)=\lambda t$）$n\to\infty$，$np$为有限数这些条件，则上式可很好地用描述泊松分布的式2—6来近似，即

$$P(r)=\frac{n!}{r!(n-r)!}(1-e^{-\lambda t})^r(e^{-\lambda t})^{n-r}$$

$$\simeq\frac{\mu^r e^{-\mu}}{r!}$$

上式中μ为在时间t内衰变掉的r个粒子数这一变量的分布的平均数也就是泊松分布的参数。