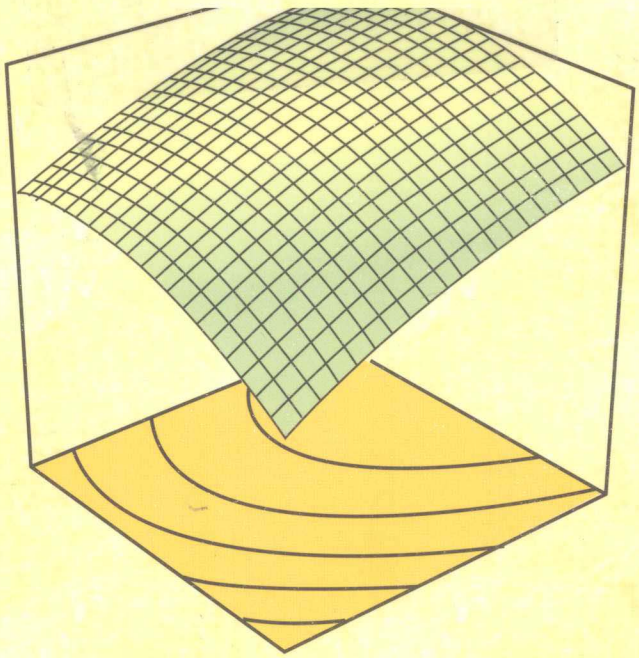
普通高等教育“十一五”规划教材

试验优化设计与统计分析

李志西　杜双奎　主编



科学出版社
www.sciencep.com

普通高等教育“十一五”规划教材

试验优化设计与统计分析

李志西　杜双奎　主编

科学出版社

北京

内 容 简 介

试验优化设计是以数理统计为基础，对试验进行优化设计与统计分析的科学方法，是科技工作者必备的基本技能。本书主要介绍了工程技术中常用的试验设计与分析方法及其在生物工程、食品工程、化学工程等技术领域中的应用。全书共分 10 章，包括试验资料的统计描述、理论分布与抽样分布、统计假设检验与参数估计、方差分析、回归与相关、试验设计基础、正交试验设计、均匀试验设计、回归试验设计、Excel 在统计分析中的应用等内容。在系统介绍常用试验设计及其统计分析方法的同时，还重点介绍了试验优化设计方法在工业生产与工程技术中的实际应用，并列举了大量实例，做到理论联系实际，便于理解和自学。深入浅出，通俗易懂，可读性强。

本书可作为轻工院校、农业院校、商学院、水产学院、粮食学院等高等院校的食品科学、食品工程、发酵工程、生物工程、食品质量与安全以及化工等专业教学用书，也可用作相关专业的成人教育教材，可供科研人员、工程技术人员、管理人员和试验工作者在学习和查阅时参考。

图书在版编目(CIP)数据

试验优化设计与统计分析/李志西，杜双奎主编. —北京：科学出版社，2010.6

（普通高等教育“十一五”规划教材）

ISBN 978-7-03-027619-3

Ⅰ.①试… Ⅱ.①李…②杜… Ⅲ.①试验设计(数学)-最佳化-高等学校-教材②试验分析(数学)-高等学校-教材 Ⅳ.①O212.6②O212.1

中国版本图书馆 CIP 数据核字(2010)第 089431 号

责任编辑：丛 楠 甄文全 陈珊珊/责任校对：张怡君

责任印制：张克忠/封面设计：耕者设计工作室

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

铁成印刷厂印刷

科学出版社发行 各地新华书店经销

*

2010 年 6 月第 一 版 开本：787×1092 1/16

2010 年 6 月第一次印刷 印张：19 3/4

印数：1—3 000 字数：465 000

定价：38.00 元

（如有印装质量问题，我社负责调换）

编委会名单

主　编　李志西（西北农林科技大学）

杜双奎（西北农林科技大学）

副主编　毛立新（山西大学）

张有林（陕西师范大学）

刘耀玺（河南科技大学）

参编者　（按姓氏笔画排列）

于修烛（西北农林科技大学）

张　辉（新疆农业大学）

张　莉（西北农林科技大学）

徐　颖（陕西科技大学）

唐明祥（石河子大学）

谢新华（河南农业大学）

目　录

第 1 章　试验资料的统计描述

1.1　常 用 术 语

1.1.1　总体与样本

根据研究目的确定的研究对象的全体称为总体（population），其中每个研究单位称为个体（individual），依据一定方法从总体中抽取的部分个体组成的集合称为样本（sample）。例如，某饮料厂某班次生产饮料 1000 瓶，则这个班次所生产的 1000 瓶饮料全体就构成研究总体，每一瓶是一个个体；从该总体中抽取 100 瓶进行分析，那么 100 瓶就为一个研究样本。含有有限个个体的总体称为有限总体（finite population）。例如，上述班次生产的饮料总体为有限总体。包含无限多个个体的总体称为无限总体（infinite population）。例如，在生物统计理论研究中服从正态分布的总体、服从 t 分布的总体，包含一切实数，属于无限总体。样本中所包含的个体数目称为样本容量或样本大小（sample size），用 n 表示。例如，上述的研究样本容量 $n=100$。通常把 $n<30$ 的样本称为小样本，$n\geqslant 30$ 的样本称为大样本。

统计分析一般是通过样本来了解总体，然而能观测到的却是样本，通过样本来推断总体是统计分析的基本特点。为了能可靠地由样本来推断总体，要求样本具有一定的含量和代表性。如何获取有代表性的样本？只有从总体中随机抽取的样本才具有代表性，这就涉及随机抽样。所谓随机抽样是指总体中每一个个体都有同等的机会被抽取而组成样本。从总体中随机抽取的部分个体所构成的样本称为随机样本。然而样本毕竟只是总体的一部分，尽管具有一定的含量，也具有代表性，但通过样本来推断总体也不可能是百分之百的正确，有很大的可靠性，也有一定的错误率。

1.1.2　参数与统计量

由总体的全部观测值计算的特征数称为参数（parameter）。参数常用希腊字母表示，如用 μ 表示总体平均数，用 σ 表示总体标准差。由样本计算的特征数称为统计量（statistic）。常用拉丁字母表示统计量，如用 $\bar{x}$ 表示样本平均数，用 S 表示样本标准差，用 R 表示极差。由于参数通常无法获得，所以总体参数常由相应的统计量来估计，如用 $\bar{x}$ 估计 μ，用 S 估计 σ 等。

1.1.3　准确性与精确性

准确性（accuracy）也称为准确度，指在调查或试验中某一试验指标的观测值与其真值接近的程度。设某一试验指标的真值为 μ，观测值为 x，若 x 与 μ 相差的绝对值 $|x-\mu|$ 越小，则观测值 x 的准确性越高；反之则越低。精确性（precision）也称为精

确度，指调查或试验中同一试验指标的重复观测值之间彼此接近的程度。若观测值彼此接近，即任意两个观测值 x_i、x_j 相差的绝对值 $|x_i-x_j|$ 越小，则观测值精确性越高；反之则越低。准确性、精确性的意义如图 1-1 所示。

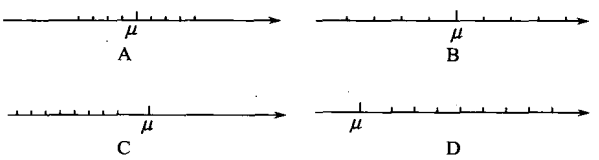


图 1-1 准确性与精确性示意图

图 1-1A 观测值密集于真值 μ 两侧，其准确性、精确性均高；图 1-1B 观测值稀疏地分布于真值 μ 两侧，其准确性高，但精确性低；图 1-1C 观测值密集于远离真值 μ 的一侧，其准确性低，但精确性高；图 1-1D 观测值稀疏地分布于远离真值 μ 的一侧，其准确性、精确性都低。

在调查或试验中应严格按照调查或试验计划进行，准确地进行观测记载，力求避免人为差错。特别要注意试验条件的一致性，除所研究的各个处理外，其他供试条件应尽量控制一致，并通过合理的调查或试验设计来提高试验的准确性和精确性。

1.1.4 随机误差与系统误差

在科学试验中，试验指标除受试验因素影响外，还会受到许多其他非试验因素干扰，从而产生误差。试验中出现的误差可分为随机误差（random error）与系统误差（systematic error）两类。随机误差也称为抽样误差（sampling error），这是由于许多无法控制的内在的和外在的偶然因素所造成的。随机误差带有偶然性，在试验中，即使十分小心也难以消除，随机误差不可避免，但可减少，随机误差影响试验的精确性。统计上的试验误差指随机误差，这种误差越小，试验的精确性越高。系统误差也称为片面误差（lopsided error），这是由于试验对象相差较大、试验周期较长、试验条件未能控制相同、测量仪器不准、标准试剂未经校正，以及观测、记载、抄录、计算中的错误所引起。系统误差影响试验的准确性，可以通过改进试验方法、正确设计试验来避免、消除。图 1-1C、D 所表示的情况，是由于出现了系统误差的缘故。一般来说，只要试验工作细致，系统误差就可以克服。图 1-1A 表示克服了系统误差的影响，且随机误差较小，因而其准确性、精确性高。

1.2 数据资料的整理

由调查或试验收集来的原始资料，往往是零乱的，无规律性可循。只有通过科学的统计整理和分析，才能发现其内部的联系和规律性。数据资料的整理是进一步统计分析

的基础。在调查或试验中，由观察、测量所得的数据按其性质的不同，一般可以分为数量资料和质量资料。

1.2.1　数据资料的分类

1. 数量资料

数量资料是指以测量、计量或计数方式获得的数据。数量资料又分为计量资料（连续性变数资料）和计数资料（间断性变数资料）两种。

1）*计量资料*　计量资料是指用测量手段得到的数据资料，即用度、量、衡等计量工具直接测定的资料，这种资料的各个观测值不一定是整数，两个相邻的整数间可以有带小数的数值出现，其小数位数的多少由度量仪器或工具的精度而定，观测值间是连续性的。因此，计量资料也称为连续性变数资料，如食品中各种营养物质的含量、苹果个体的重量等。

2）*计数资料*　计数资料指用计数方式得到的数据资料。在这类资料中，各个观测值只能用整数表示，在两个相邻整数间不可能有带小数的数值出现，各观测值是不连续的。因此该类资料也称为不连续性变数资料或间断性变数资料，如一箱饮料的瓶数、微生物的个数等。

2. 质量资料

质量资料是指能观察到而不能直接测量的，只能用文字来描述其特征而获得的资料，如食品颜色、风味、酒的风格等。这类资料本身不能直接用数值表示，要获得这类数据资料，需对其观测结果作数量化处理，其方法有统计次数法和评分法等。

1）*统计次数法*　在一定的总体或样本中，根据某一质量性状的类别统计其次数，以次数作为质量性状的数据。例如，在研究批次产品合格数与次品数时，可以统计其合格与次品个数。这种由质量性状数量化得来的资料又称为次数资料。

2）*评分法*　对某一质量性状，因其类别不同，分别给予评分。例如，分析面包的质量时，可以按照国际面包评分细则进行打分，综合评价面包质量。

1.2.2　数据资料的整理

根据资料中观测值的多少确定是否分组。当观测值较少（$n \leqslant 30$）时，不必分组，可直接进行统计分析。当观测值较多（$n > 30$）时，宜将观测值分成若干组，以便统计分析。将观测值分组后，制成次数分布表，即可得到资料的集中和变异情况。不同类型的资料，其整理方法略有不同。

1. 连续性变数资料的整理

连续性资料的整理，需要先确定全距、组数、组距、组中值以及组限，然后将全部观测值计数归组。下面以 100 听罐头的净重资料为例来说明其整理的方法及步骤。

例 1-1　为分析某食品厂的罐头产品质量，随机抽取 100 听罐头样品，其净重测定结果见表 1-1，试整理成次数分布表。

表 1-1 100 听罐头样品的净重（g）

342.1	340.7	348.4	346.0	343.4	342.7	346.0	341.1	344.0	348.0
346.3	346.0	340.3	344.2	342.2	344.1	345.0	340.5	344.2	344.0
343.5	344.2	342.6	343.7	345.5	339.3	350.2	337.3	345.3	358.2
344.2	345.8	331.2	342.1	342.4	340.5	350.0	343.2	347.0	340.2
344.0	353.3	340.2	336.3	348.9	340.2	356.1	346.0	345.6	346.2
340.6	339.7	342.3	352.8	342.6	350.3	348.5	344.0	350.0	335.1
340.3	338.2	345.5	345.6	349.0	336.7	342.0	338.4	343.9	343.7
341.1	347.1	342.5	350.0	343.5	345.6	345.0	348.6	344.2	341.1
346.8	350.2	339.9	346.6	339.9	344.3	346.2	338.0	341.1	347.3
347.2	339.8	344.4	347.2	341.0	341.0	343.3	342.3	339.5	343.0

（1）求全距。全距是资料中最大值与最小值之差，又称为极差（range），用 R 表示，即

$$R = max(x_i) - min(x_i)$$

式中，x_i 为观测值。

表 1-1 中，罐头样品最大净重为 358.2 g，最小净重为 331.2 g，因此

$$R = 358.2 - 331.2 = 27.0\ \text{g}$$

表 1-2 样本含量与组数

样本含量（n）	组 数
60～100	7～10
100～200	9～12
200～500	12～17
500 以上	17～30

（2）确定组数。组数的多少要根据样本含量及资料的变动范围大小而定，一般以既简化资料又能反映资料的规律性为原则。组数要适当，不宜过多，也不宜过少。分组越多所求得的统计量越精确，但增大了运算量；若分组过少，资料的规律性不明显，计算出的统计量的精确性也较差。一般组数的确定见表 1-2。

在本例中，n=100，根据表 1-2，确定组数为 9 组。

（3）确定组距。每组最大值与最小值之差称为组距（class interval），记为 i。分组时要求各组的组距相等。组距的大小由全距与组数确定，即

$$i = \frac{\text{全距}}{\text{组数}}$$

本例组距 i=27.0/9=3.0。

（4）确定组限及组中值。各组的最大值与最小值称为组限（class limit）。最小值称为下限（lower limit），最大值称为上限（upper limit）。每一组的中点值称为组中值（class value），它是该组的代表值。所以，组中值与组限、组距的关系如下：

$$\text{组中值} = \frac{\text{组下限} + \text{组上限}}{2} = \text{组下限} + \frac{\text{组距}}{2} = \text{组上限} - \frac{\text{组距}}{2}$$

由于相邻两组的组中值的距离等于组距，所以当第一组的组中值确定以后，加上

组距就是第二组的组中值，第二组的组中值加上组距就是第三组的组中值，依此类推。

组距确定后，首先要选定第一组的组中值。在分组时为了避免第一组中观察值过多，一般第一组的组中值以接近或等于资料中的最小值为好。第一组组中值确定后，该组组限即可确定，其余各组的组中值和组限也可相继确定。注意，最末一组的上限应大于资料中的最大值。

如例 1-1 中，最小值为 331.2，第一组的组中值可取 331.0，因组距为 3.0，因此，第一组的下限为

$$331.0-\frac{3.0}{2}=329.5$$

第一组的上限也就是第二组的下限，应为

$$329.5+3.0=332.5$$

第二组的上限也就是第三组的下限，应为

$$332.5+3.0=335.5,\cdots$$

依此类推，分组为 329.5～332.5，332.5～335.5，…

通常将上限略去，如第一组记为 329.5～ ，第二组记为 332.5～……

(5) 归组计数，制作次数分布表。将资料中的每一个观测值逐一归组，统计每组内所包含的观测值个数，制作次数分布表。一般将正好等于前一组上限和后一组下限的数据归入后一组。

次数分布表不仅便于观察资料的规律性，而且可根据它绘成次数分布图并计算平均数、标准差等统计量。表 1-3 为 100 听罐头净重的次数分布表。

表 1-3　100 听罐头净重的次数分布

组限	组中值 (x)	次数 (f)
329.5～	331.0	1
332.5～	334.0	1
335.5～	337.0	6
338.5～	340.0	21
341.5～	343.0	32
344.5～	346.0	23
347.5～	349.0	12
350.5～	352.0	2
353.5～	355.0	1
356.5～	358.0	1

从表 1-3 中可以看出，100 听罐头的单听净重多数为 343.0 g，约占观测值总个数的 1/3，用它来描述罐头单听净重的平均水平，有较强的代表性。每听罐头净重小于 332.5g 及大于 356.5g 的为极少数。100 听罐头净重分布基本以 343.0g 为中心，向两边

做递减对称分布。

(6) 次数分布图。次数分布一般用次数分布图表示。次数分布图主要有直方图、折线图两种。次数分布图以分组组中值为横坐标、次数为纵坐标绘制。如图 1-2、图 1-3 所示，由次数分布图明显看出 100 听罐头的净重分布情况以及平均净重。

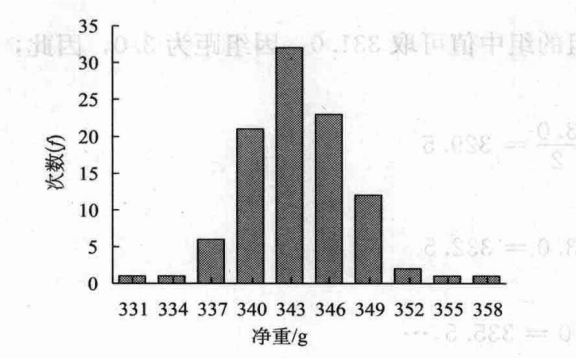


图 1-2 100 听罐头净重次数分布直方图

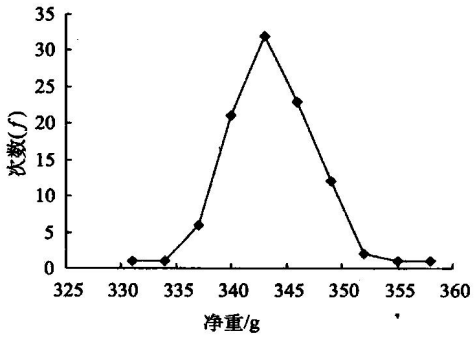


图 1-3 100 听罐头净重次数分布折线图

2. 间断性变数资料的整理

例 1-2 以 50 盒鲜枣不合格枣数为资料来说明间断性资料的整理分析（表 1-4）。

表 1-4 50 盒鲜枣每盒检出不合格枣数

21	20	20	21	23	22	22	22	21	22
24	22	19	22	21	21	21	22	22	24
21	22	22	23	22	23	22	22	22	23
23	22	23	22	19	22	23	20	22	21
21	21	22	22	23	22	22	22	23	22

表 1-5 50 盒鲜枣不合格枣数次数分布表

不合格数	次数（f）
19	2
20	3
21	10
22	24
23	9
24	2
合计	50

有些计数资料，观察值较多，变异范围较大，若以每一观察值为一组，则组数太多，而每组内包含的观察值太少，资料的规律性不明显。对于这样的资料，可扩大为以几个相邻观察值为一组，适当减少组数，这样资料的规律性较明显，对资料做进一步计算分析也比较方便。

从表 1-5 可以看出，86%盒的不合格枣数为 21～23 颗。仅有 8%盒的不合格枣数小于 19 颗或大于 24 颗。

1.3　变数的统计描述

1.3.1　统计特征数

通过数据资料的整理，得到次数分布表和次数分布图，可形象、直观地表示出资料的集中性和离散性。所谓集中性就是变数在趋势上向某一中心聚集，或以某一数值为中心而分布的性质。所谓离散性则是变数在趋势上分散离中的变异性质。

度量集中性（分布中心点）的统计特征数是平均数，包括算术平均数（arithmetic mean）、几何平均数（geometric mean）、调和平均数（harmonic mean）、中位数（median）和众数（mode）等。度量离散性（分布范围）的统计特征数一般有极差（range）、方差和标准差（standard deviation）等。

1.3.2　平均数

平均数（mean）是统计学中最常用的特征数，用来描述资料的集中性，即指资料中数据集中较多的中心位置。平均数可以分为算术平均数、几何平均数、调和平均数等。其中最常用的是算术平均数，简称平均数。

1. 算术平均数

将观测值的总和除以观测值个数所得的商称为算术平均数，记为 $\bar{x}$。算术平均数可根据样本大小及分组情况而采用直接法或加权法计算。

1）*直接法*　　直接法主要用于样本含量 $n\leqslant 30$ 且未经分组资料平均数的计算。

设 $\bar{x}$ 代表 x_1，x_2，…，x_n 等 n 个变数的算术平均数，则有如下关系式：

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

式中，$\sum$ 为加和符号（summation）；$\sum_{i=1}^{n}x_i$ 为 $x_1+x_2+\cdots+x_n$ 的总和。在计算意义明确时，$\sum_{i=1}^{n}x_i$ 可以简写成 $\sum x$。若欲求 2，4，6，8，15 的算术平均数，其计算方法如下：

$$\bar{x}=\frac{2+4+6+8+15}{5}=7$$

例 1-3　对某乳品厂生产的 10 袋小包装产品净重进行测定，结果为 50.0g、52.0g、53.5g、56.0g、58.5g、60.0g、48.0g、51.0g、50.5g、49.0g，求其平均数。

由于

$$\begin{aligned}\sum x&=50.0+52.0+53.5+56.0+58.5+60.0+48.0+51.0+50.5+49.0\\&=528.5\\n&=10\end{aligned}$$

所以，算术平均数

$$\bar{x}=\frac{528.5}{10}=52.85\text{g}$$

小包装产品的净重平均为 52.85g。

2）*加权法*　对于样本含量 $n \geqslant 30$ 且已分组的资料，可在次数分布表的基础上采用加权法计算平均数，计算公式为

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \cdots + f_kx_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k}(f_ix_i)}{\sum_{i=1}^{k} f_i} = \frac{\sum(fx)}{\sum f} \tag{1-2}$$

式中，x_i 为第 i 组的组中值；f_i 为第 i 组的次数；k 为分组数。

由式（1-2）计算的平均数称为加权平均数（weighted mean）。f_i 是变量 x_i 所具有的“权”，即变量 x_i 所占的比重。因为

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + \cdots + f_k = n$$

故

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k}(f_ix_i) = \frac{1}{n}\sum(fx)$$

根据表 1-1 资料计算 100 听罐头每听净重的加权平均数 $\bar{x}$，则

$$\bar{x} = \frac{\sum(fx)}{\sum f} = \frac{(331.0 \times 1 + 334.0 \times 1 + 337.0 \times 6 + \cdots + 358.0 \times 1)}{100} = 343.66\text{g}$$

100 听罐头每听净重的加权平均数为 343.66g。

在试验统计中，习惯上将样本平均数用 $\bar{x}$ 表示，总体平均数用 μ 表示。

2. 几何平均数

当遇到计算平均增长率时，常常以几何平均数表示其平均值。设 G 为 x_1，x_2，…，x_n 这 n 个数据的几何平均数，则有

$$G = (x_1 \cdot x_2 \cdot x_3 \cdot \cdots \cdot x_n)^{\frac{1}{n}} \tag{1-3}$$

对式（1-3）取对数：

$$\lg G = \frac{\lg x_1 + \lg x_2 + \cdots + \lg x_n}{n} = \frac{\sum \lg x}{n} \tag{1-4}$$

可见，几何平均数是变量对数的算术平均数的反对数值。

如果是一个分组数列时，则几何平均数可用下列公式计算：

$$G = (x_1^{f_1} \cdot x_2^{f_2} \cdot \cdots \cdot x_n^{f_n})^{\frac{1}{n}} \tag{1-5}$$

两边取对数，则 $\lg G = \frac{1}{n}\sum_{i=1}(f_i \lg x_i)$，其中 $n = \sum_{i=1}^{n} f_i$。

由式（1-5）计算的几何平均数称为加权几何平均数。

3. 调和平均数

计算平均速率时需用调和平均数，用 H 表示。

通常，先计算倒数平均数 $\frac{1}{H}$，即

$$\frac{1}{H}=\frac{1}{n}\sum\frac{1}{x} \tag{1-6}$$

那么

$$H=\frac{n}{\sum\frac{1}{x}} \tag{1-7}$$

可见，调和平均数是变量倒数的算术平均数的倒数。

4. 中位数

中位数是指资料中的观测值由大到小依次排列后，居于中间位置的那个观测值。中位数也称中数，记作 M_d。

中位数的计算比较方便。将资料数据按大小顺序排列，中位数在数列中的位次可用算式（$n+1$）/2 来确定，处于这一位次的数就是中位数 M_d。如果资料中数据个数为偶数时，则其中间两个数的算术平均数为中位数。

例 1-4　某食品生产小组共有 5 名工人，他们日生产的产品件数的排列次序为：18，20，23，25，26。则中位数的位次为（$n+1$）/2=（5+1）/2=3，即第 3 名工人所完成的件数（23 件）是中位数。若现有第 6 名工人加入该小组，其日生产件数为 29 件，则（$n+1$）/2=(6+1) /2=3.5，说明中位数在第 3 名和第 4 名工人所完成的件数之间，中位数取值为（23+25）/2=24 件。

如果是分组数列求中位数，可用下式计算：

$$M_d=x_0+\frac{\frac{\sum f}{2}-S_{m-1}}{f_m}\cdot d \tag{1-8}$$

式中，M_d 为中位数；$\sum f$ 为次数总和；S_{m-1}为累积到中位数前一组的次数之和；f_m 为中位数的次数；d 为中位数的组距数值；x_0 为次数最多的组的下限。

5. 众数

众数（M_0）是变异数列中出现次数最多的那个数值，在频率分布图中，就是频率具有最大值所对应的那个变数值。众数也表示数据集中的趋向。在非对称的频率分布中，平均数、中位数和众数三者并不重合。频率曲线越不对称，三者差别就越大。

1.3.3　标准差和变异系数

1. 标准差的意义

用平均数作为样本的代表，其代表性强弱受样本资料中各观测值变异程度的影响。如果各观测值变异小，则平均数对样本的代表性强；如果各观测值变异大，则平均数对样本的代表性弱。因而仅用平均数对一个资料的特征作统计描述是不全面的，还需引入一个表示资料中观测值变异程度大小的统计量。

全距（极差）是表示资料中各观测值变异程度大小最简便的统计量。全距大，则资料中各观测值变异程度大；全距小，则资料中各观测值变异程度小。但是全距只利用了资料中的最大值和最小值，并不能准确表达资料中各观测值的变异程度，比

较粗略。当资料很多而又要迅速对资料的变异程度作出判断时，可以利用全距这个统计量。标准差广泛用来度量样本各观测值间的变异程度和平均数的代表情况。标准差充分考虑了各个变数与平均数的离差。每个变数与平均数相差越小，则样本变异程度越小；反之越大。如果每个变数与平均数之差为零，这表示每个变数与平均数没有差异。标准差是用各变数与平均数差的大小来度量变异程度的一个统计量。

2. 标准差的计算

为了准确表示样本内各个观测值的变异程度，人们首先会考虑以平均数为标准，求出各个观测值与平均数的离差，即 $(x-\bar{x})$，称为离均差。虽然离均差能表达一个观测值偏离平均数的性质和程度，但因为离均差有正、有负，离均差之和为零，即 $\sum(x-\bar{x})=0$，因而不能用离均差之和，即 $\sum(x-\bar{x})$ 来表示资料中所有观测值的总偏离程度。为了合理地计算出平均差异，我们先将各个离均差平方，即 $(x-\bar{x})^2$，再求离均差平方和，即 $\sum(x-\bar{x})^2$，简称平方和，记为 SS。

由于离差平方和常随样本大小而改变，为消除样本大小的影响，用平方和除以样本大小，即 $\sum(x-\bar{x})^2/n$。为了使所得的统计量是相应总体参数的无偏估计量，统计学证明，在求离均差平方和的平均数时，分母为自由度 $n-1$。于是，我们采用统计量 $\sum(x-\bar{x})^2/(n-1)$ 表示资料的变异程度。统计量 $\sum(x-\bar{x})^2/(n-1)$ 称为均方(mean square，MS)，又称样本方差，记为 S^2，即

$$S^2=\frac{\sum(x-\bar{x})^2}{n-1} \tag{1-9}$$

相应的总体参数称为总体方差，记为 σ^2。对于有限总体而言，σ^2 的计算公式为

$$\sigma^2=\frac{\sum(x-\mu)^2}{N} \tag{1-10}$$

统计学上把 S^2 的平方根称为标准差(standard deviation)，记为 S，其单位与观测值度量单位相同。由样本资料计算标准差的定义公式为

$$S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \tag{1-11}$$

这里 $n-1$ 为自由度，应用自由度的目的是为了减少抽样误差的影响。例如，一个样本含有 n 个变数，从理论上说，n 个变数与 $\bar{x}$ 之差得到 n 个离均差，但是，其中 $n-1$ 个是可以自由变动的，最后一个离均差受 $\sum(x-\bar{x})=0$ 这一条件的限制而不得自由变动。例如，有 5 个变数，其中有 4 个离均差为 -2、-1、1、2，则第 5 个离均差值必须等于 0。如果其中前 4 个的离均差为 -1、0、1、2，则第 5 个离均差只能等于 2，这样才能使得离均差的总和等于 0。所以，这 5 个离均差中，只有 4 个能自由变动，自由度就是 $n-1=4$。通常，自由度等于样本变数的总个数减去计算过程中使用的条件数。在计算标准差时，条件只有一个，即 $\sum(x-\bar{x})=0$，故自由度为 $n-1$。如果计算其他统计数时，要是应用 2 个条件，自由度就为 $n-2$，应用 k 个条件，则自由度为 $n-k$。