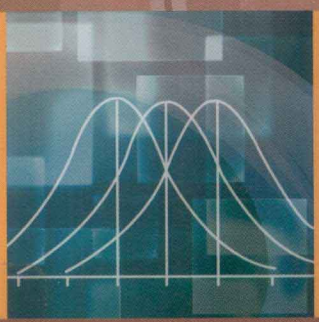
GAODENG XUEXIAO ZHUANYE JIAOCAI

· 高等学校专业教材 ·

食品试验优化设计

杜双奎　李志西　主编

FOOD EXPERIMENTAL DESIGN AND STATISTICAL ANALYSIS



中国轻工业出版社

高等学校专业教材

食品试验优化设计

杜双奎　李志西　主　编
蔡健荣　主　审

中国轻工业出版社

图书在版编目（CIP）数据

食品试验优化设计/杜双奎，李志西主编. —北京：
中国轻工业出版社，2011.8
高等学校专业教材
ISBN 978-7-5019-8276-9

Ⅰ.①食… Ⅱ.①杜… ②李… Ⅲ.①食品工业－试
验设计－高等学校－教材
Ⅳ. ①TS2－33

中国版本图书馆 CIP 数据核字（2011）第 101862 号

责任编辑：张 靓 责任终审：简延荣 封面设计：锋尚设计
版式设计：宋振全 责任校对：杨 琳 责任监印：张 可

出版发行：中国轻工业出版社（北京东长安街 6 号，邮编：100740）
印 刷：北京君升印刷有限公司
经 销：各地新华书店
版 次：2011 年 8 月第 1 版第 1 次印刷
开 本：787×1092 1/16 印张：20.5
字 数：498 千字
书 号：ISBN 978-7-5019-8276-9 定价：38.00 元
邮购电话：010-65241695 传真：65128352
发行电话：010-85119835 85119793 传真：85113293
网 址：http://www.chlip.com.cn
Email:club@chlip.com.cn
如发现图书残缺请直接与我社邮购联系调换
100083J1X101ZBW

前　言

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。

本书在保持本学科的系统性和科学性的前提下，注意引入学科发展的新知识、新成果，注重拓宽学生的知识面和提高实践能力以及统计分析与计算机科学的结合，力求体现强基础、重应用和当前进行的素质教育和创新教育的教学目标。

本书主要介绍了工程研究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。全书共分八章，由西北农林科技大学杜双奎、李志西担任主编，江苏大学蔡健荣教授担任主审。参加编写的人员分工如下：第一章由李志西、张华编写，第二章由乐素菊编写，第三章由赵胜娟编写，第四章由程江峰编写，第五章由林颢编写，第六章由艾对元编写，第七章由杜双奎编写，第八章由陈全胜编写，附表由于修烛、王鑫组织，全书由杜双奎统稿。本书在系统介绍常用试验设计及其统计分析方法的同时，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，列举了大量实例，做到理论与实际的联系，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、工程技术人员和试验工作者学习和查阅。

在编写大纲修订与完善过程中，江苏大学蔡健荣教授提出了宝贵意见，在此表示诚挚谢意。编写中引用和参考了有关中外文献和专著，编者对这些文献和专著的作者、对大力支持编写工作的中国轻工业出版社一并表示衷心的感谢！在编写过程中，也得到了各参编院校有关领导及其他有关方面的大力支持，谨此致谢。

由于编写人员水平有限，书中错误、缺点在所难免，敬请广大读者批评指正，以便修订、补充和完善。

编　者

目　录

第一章　试验设计与数理统计基础

第一节　统计常用术语

一、总体与样本

在数理统计中，根据研究目的确定的研究对象的全体集合称为总体（population），其中每一研究单位（元）称为个体（individual）；依据统计原理由总体中抽取的部分个体组成的集合称为样本（sample），它是测定、分析、研究的直接对象。例如某方便面企业的质检部门为检测某班次当天生产的盒装方便面质量，从中随机抽取 50 份进行分析检测，那么这个班次当天生产的所有盒装方便面就是质检的研究总体，每 1 份盒装方便面就是一个个体，质检人员随机抽取的 50 份就是一个研究样本。含有有限个个体的总体称为有限总体（finite population），含有无限个个体的总体称为无限总体（infinite population）。样本中所包含的个体数目称为样本容量或大小（sample size），记为 n。通常 $n<30$ 的样本为小样本，$n\geqslant 30$ 的样本为大样本。

二、参数与统计量

用来描述总体特征的量称为参数（parameter）。常用希腊字母表示，如用 μ 表示总体平均数，用 σ^2 表示总体方差，用 σ 表示总体标准差。

用来描述样本特征的量称为统计量（statistic）或统计数。常用拉丁字母表示，例如用 $\bar{x}$ 表示样本平均数，用 S^2 表示样本方差，用 S 表示样本标准差。总体参数通常是无法获得的，常由相应的统计量来估计，例如用 $\bar{x}$ 估计 μ，用 S^2 估计 σ^2 等。

三、准确性与精确性

准确性（accuracy）也称准确度，是指试验中某一指标或性状的观测值与其真值接近的程度。假设某一指标或性状的真值为 μ，观测值为 x，那么绝对值 $|x-\mu|$ 越小，表明观测值 x 的准确性越高；反之越低。

精确性（precision）也称精确度，是指同一指标在重复试验中，其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；反之越低。准确性、精确性的意义如图 1-1 所示。

假如试验理论真值 μ 在同心圆的中心，那么，图 1-1（a）观测值密集于真值 μ 附近，其准确性高，精确性亦高；图 1-1（b）观测值较稀疏地分布于真值 μ 周围，其准确性高，但精确性低；图 1-1（c）观测值密集于真值 μ 的一侧，但远离真值 μ，准确性低，精确性高；图 1-1（d）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。

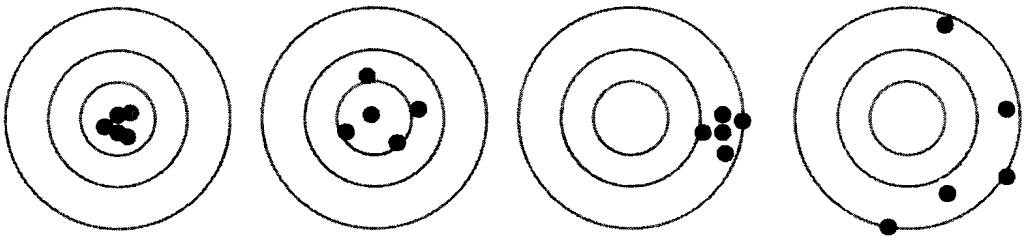


图 1－1　准确性与精确性示意图

四、随机误差与系统误差

在科学试验中，试验结果除受试验因素影响外，还会受到许多其它非试验因素的干扰，从而产生误差。试验误差有随机误差（random error）与系统误差（systematic error）之分。

随机误差也称抽样误差（sampling error），这是由于许多无法控制的内在和外在的偶然因素所造成。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。随机误差影响试验结果的精确性。统计上提到的试验误差通常指随机误差，这种误差愈小，试验的精确性愈高。

系统误差也称片面误差（lopsided error），这是由于试验对象相差较大，或试验周期较长，试验条件控制不一致，或测量仪器不准，或标准试剂未经校正，以及观测、记载、抄录、计算中的错误所引起。系统误差影响试验结果的准确性，但可以通过改进试验方法和试验设计方案来避免或消除。图 1－1（c）所表示的情况就是由于出现了系统误差的缘故。通常，只要试验工作做得精细，系统误差就可以克服。图 1－1（a）为理想的试验结果，准确性高，精确性也高，这是克服了系统误差、降低随机误差而获得的。

第二节　统计特征数

一、平　均　数

平均数是度量数据资料集中性的统计特征数，有算术平均数、几何平均数、调和平均数、中位数和众数等。其中最常用的是算术平均数，简称平均数。

1. 算术平均数（arithmetic mean）

观测值的总和除以观测值个数所得数值称为算术平均数，记为 $\bar{x}$。

假设 $\bar{x}$ 为 x_1，x_2，…，x_n 等 n 个观测值的算术平均数，则

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

上式中 $\sum$ 为加和符号；$\sum_{i=1}^{n} x_i$ 表示由第 1 个观测值 x_1 累加到第 n 个观测值 x_n，即 x_1 +

$x_2 + \cdots + x_n$的总和。在计算意义明确时，$\sum_{i=1}^{n} x_i$ 可简写成 $\sum x$ 。

对于样本资料数据较多的，也可在次数分布的基础上采用加权法计算平均数：

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

式中　x_i——第 i 组的组中值

f_i——第 i 组的次数

k——分组数

由上式计算的平均数也称为加权平均数（weighted mean）。f_i 是变量 x_i 所具有的“权”。由于 $\sum_{i=1}^{k} f_i = f_1 + f_2 + \cdots + f_k = n$，故 $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_ix_i = \frac{1}{n}\sum fx$ 。

2. 几何平均数（geometric mean）

在统计分析中，当资料中的观测值呈几何级数变化趋势，需要计算平均增长率时，常以几何平均数表示其平均值，以 G 标记。

假设 G 为 x_1，x_2，…，x_n　n 个数据的几何平均数，则：

$$G = (x_1x_2x_3\cdots x_n)^{\frac{1}{n}} \tag{1-3}$$

对式（1-3）取对数，得：

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

由式（1-4）可以看出，几何平均数是观测值对数的算术平均数的反对数值。

如果研究仅有最初观测值 a_1 和最末水平观测值 a_n 时，则其几何平均数为：

$$\dot{G} = \sqrt[n-1]{\frac{a_n}{a_1}} = \left(\frac{a_n}{a_1}\right)^{\frac{1}{n-1}} \tag{1-5}$$

式中　n——数列中的项数

3. 调和平均数（harmonic mean）

计算平均速率时需用调和平均数，用 H 表示。

若 H 为 x_1，x_2，…，x_n 的调和平均数，那么，

$$H = \frac{n}{\sum\left(\frac{1}{x}\right)} \tag{1-6}$$

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（median）

中位数是指资料中的观测值由大到小（或由小到大）依次排列后，居于中间位置的那个观测值。中位数也称为中数，记作 M_d。

当观测值 n 为偶数时，则第 $\frac{n}{2}$ 与第 $\frac{n}{2}+1$ 两个观测值的平均数为中位数。当观测值的个数 n 为奇数时，中位数的位次可用 $\frac{n+1}{2}$ 来确定，即 $x_{(n+1)/2}$ 为中位数。

5. 众数（mode）

众数是指数据资料中次数出现最多的那个数值，记作 M_0。在非对称的资料数据分布中，平均数 $\bar{x}$、中位数 M_d和众数 M_0三者并不重合。平均数、中位数和众数均可反映数据的集中性，在实际中应根据具体情况而选择应用。平均数简明易懂，便于运算，因此使用最多。但当数据资料有异常大（小）值时，平均数易受其影响，失去代表性。这时，常考虑使用中位数 M_d，而众数 M_0在市场销售中会用到。

二、变 异 数

度量数据离散性（分布范围）的统计特征数称为变异数，通常有极差、方差、标准差和变异系数等。

1. 极差（range）

极差是数据资料中最大值与最小值之差，表示资料中各观测值离散程度大小最简便的统计量，记为 R。

$$R = \max(x_1, x_2, \cdots, x_n) - \min(x_1, x_2, \cdots, x_n) \tag{1-7}$$

极差越大，表明数据资料中的观测值离散程度越大，极差越小，观测值的离散程度越小。但是极差仅仅利用了资料中的最大值和最小值的信息，并不能准确表达出数据资料中各个观测值的变异程度，比较粗。当资料很多而又要迅速对资料的离散程度作出判断时，可利用极差这一统计量。

2. 方差（variance）

为了能够准确地表示样本内各个观测值的变异程度，我们以平均数为标准，求出各个观测值与平均数的离差，即 $(x-\bar{x})$，称为离均差。离均差能表达每一个观测值偏离平均数的性质和程度，但离均差之和为零。为了合理地计算出平均差异，我们可采用将离均差平方的办法来解决离均差之和为零的问题。即先将各个观测值的离均差平方，即 $(x-\bar{x})^2$，再将离均差平方加和，求其总离均差平方和，即 $\sum (x-\bar{x})^2$，也称为偏差平方和，简称平方和，记为 SS。

为了消除样本大小对离均差平方和的影响，可用平方和除以样本大小，即 $\sum (x-\bar{x})^2/n$，称为平均离均差平方和。统计学证明，$\sum (x-\bar{x})^2/(n-1)$ 是相应总体方差（σ^2）的无偏估计值，可以度量资料的变异程度。所以，统计量 $\sum (x-\bar{x})^2/(n-1)$ 称为均方（mean square 缩写为 MS），也称样本方差，记为 S^2，即

$$S^2 = \frac{\sum (x-\bar{x})^2}{(n-1)} \tag{1-8}$$

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，

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} \tag{1-9}$$

3. 标准差（standard deviation）

统计学上把方差 S^2 的正平方根值称为标准差，记为 S，其单位与观测值的度量单位相同。由样本资料计算标准差的定义公式为：

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \tag{1-10}$$

$n-1$ 在统计上称为自由度，是指独立观测值的个数，用 df 表示。其统计意义是指样本内独立而能自由变动的离均差个数。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，所以自由度为 $n-1$。若计算其它统计量时，如果受到 k 个条件的限制，则其自由度为 $n-k$。若样本容量很大时，可不用自由度，直接用 n 亦可。

由于
$$\sum (x - \bar{x})^2 = \sum (x^2 - 2x\bar{x} + \bar{x}^2) = \sum x^2 - 2\bar{x}\sum x + n\bar{x}^2$$
$$= \sum x^2 - 2\frac{\left(\sum x\right)^2}{n} + n\left(\frac{\sum x}{n}\right)^2$$
$$= \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

所以，

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

相应的总体参数称为总体标准差，记为 σ。对于有限总体，σ 可根据下列公式计算：

$$\sigma = \sqrt{\sum (x - \mu)^2 / N} \tag{1-12}$$

在统计学中，总体标准差 σ 常用样本标准差 S 来估计。

4. 变异系数（coefficient of variation）

变异系数是指标准差相对于平均数的百分数，用符号 CV 表示，是衡量资料中各观测值变异程度的另一个统计量，也称相对标准偏差。当两个或多个资料变异程度相互比较时，如果度量单位和（或）平均数不同，需采用标准差与平均数的比值（相对值）来比较，这个比值称为变异系数。变异系数可以消除单位和（或）平均数不同对两个或多个资料变异程度比较的影响，用 CV 可以比较不同样本相对变异程度的大小。

变异系数的计算公式为：

$$CV = \frac{S}{\bar{x}} \times 100\% \tag{1-13}$$

式中　CV——变异系数

S——标准差

$\bar{x}$——平均数

三、平均数和标准差的性质

1. 平均数的性质

（1）变量 x 的各观测值与其平均数 $\bar{x}$ 之差的总和（离均差和）等于零，即 $\sum_{i=1}^{n} (x - \bar{x}) = 0$。

（2）样本中各个观察值与其平均数离差平方和为最小，即离均差平方和最小。设 a

为任意常数，则有 $\sum (x-a)^2 \geqslant \sum (x-\bar{x})^2$ 。

2. 标准差的性质

（1）标准差的大小受每个观测值的影响，如观测值间变异大，其离均差亦大，由此求得的标准差必然大；反之则小。

（2）变量 x 各个观察值加上或减去一个常数 a，其标准差不变，常数的标准差等于零。

（3）当样本资料中每个观察值乘以或除以一个不等于零的常数 a 时，所得的标准差是原标准差的 a 倍或 $1/a$。

（4）标准差主要用以衡量资料的变异程度，评定平均数的代表性，估计资料中变量的分布情况。当数据资料为正态分布时，以平均数为中心左右各取 1 个标准差，即 $\bar{x} \pm S$ 范围内可包括全部数据的 68.27%；$\bar{x} \pm 2S$ 范围内可包括全部数据的 95.46%，$\bar{x} \pm 3S$ 范围内可包括全部数据的 99.73%。

第三节　试验数据的分类与整理

一、试验数据分类

1. 数量资料

数量资料是指通过测量、计量或计数方式而获得的数据，有计量资料（连续性变数资料）和计数资料（间断性变数资料）之分。

（1）计量资料　指用度、量、衡等计量工具直接测定而获得的数据资料。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，各个观测值间的变异是连续性的。因此，计量资料又称为连续性变异资料。如食品中各种营养成分的含量、苹果个体的重量、小麦中淀粉的含量等。

（2）计数资料　指用计数方式得到的数据资料。在这类资料中，各个观测值只能以整数表示，各个观测值不是连续的，因此该类资料也称为不连续性变异资料或间断性变异资料。如盒装方便面的份数、一箱饮料的瓶数、微生物的个数、腐烂果品的个数等。

2. 质量资料

质量资料是指不方便直接测量，只能通过观察，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。这类特征不能直接用数值表示，要获得这类特征的数据资料，需对其观察结果作必要的数量化处理。

（1）评分法　这是食品感官评价中常用的一种方法。一般请若干有经验的人，根据相关评判标准，对试验产品的指标综合评判打分，用评分进行统计分析。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。

（2）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次数资料。

（3）分级法　将变异的性状分成几级，每一级别指定以适当的数值表示。例如食品褐变程度按深浅分为五级，由这种方法所得到的数据类似于间断性资料。

（4）秩次法　将各种处理按指标性状的好坏依次排队，排队的顺序为秩，用处理的秩和进行统计分析，这在食品感官评定过程中常用到。

（5）化学分析法　对于某些质量指标，虽然用分级法、评分法、统计次数法也能得到数量资料，但得到的多数是次数资料。若借助化学分析手段即可得到计量资料。例如果汁的色泽可通过测定果汁中花青苷的光密度来表示，澄清度可用测定其透光率来表示等。这种资料属于计量资料，易于分析。

除以上几种方法以外，也可以借助必要的先进仪器来评价质量指标，获得数量资料。如质构仪、色差计、色谱仪、质谱仪等。

二、试验数据整理

当资料观测值较少（$n \leqslant 30$）时，不必分组，可直接进行统计分析。当观测值较多（$n > 30$）时，需将观测值分成若干组，以便统计分析。

[例 1-1]　国家质检部门对某企业生产的小包装豆粉净重量进行抽检，随机抽取100份样品，其测定结果见表1-1，试整理分析。

表 1-1　　100 份样品的净重　　单位：g

49.8	49.7	50.4	50.2	49.9	49.9	50.2	49.7	50.0	50.4
50.2	50.2	49.6	50.0	49.8	50.0	50.1	49.7	50.0	50.0
50.0	50.0	49.9	50.0	50.2	49.5	50.6	49.3	50.1	51.4
50.0	50.2	48.7	49.8	49.8	49.7	50.6	49.9	50.3	49.6
50.0	50.9	49.6	49.2	50.5	49.6	51.2	50.2	50.2	50.2
49.7	49.6	49.8	50.9	49.9	50.6	50.5	50.0	50.6	49.1
49.6	49.4	50.2	50.2	50.5	49.3	49.8	49.4	50.0	50.0
49.7	50.3	49.9	50.6	50.0	50.2	50.1	50.5	50.0	49.7
50.3	50.6	49.6	50.3	49.6	50.0	50.2	49.4	49.7	50.3
50.3	49.6	50.0	50.3	49.7	49.7	49.9	49.8	49.6	49.9

1. 求全距

全距是资料中最大值与最小值之差，又称极差（range），用 R 表示，即

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

表1-1中，100份样品最大净重值为51.4，最小净重值为48.7，因此

$$R = 51.4 - 48.7 = 2.7$$

2. 确定组数与组距

组数的多少视样本大小而定，一般以达到既简化资料又不影响反映资料的规律性为原则。一般组数的确定，可参考表1-2。

表 1-2　　样本容量与组数

样本容量 n	组　数
10～100	7～10
100～200	9～12
200～500	12～17
500 以上	17～30

对本例而言，样本容量 $n=100$，根据表 1－2 初步确定组数为 9 组。

组距 i 由全距与组数计算

$$\text{组距 } i = \text{全距}/\text{组数}$$

本例 $i=2.7/9=0.3$，即每组最大值与最小值之差为 0.3。

3. 确定组限、组中值

每一组中的最小值称为下限，最大值称为上限，中间值称为组中值，它是该组的代表值。组中值与组限、组距的关系为：

$$\text{组中值} = \frac{\text{组下限} + \text{组上限}}{2} = \text{组下限} + \frac{1}{2}\text{组距} = \text{组上限} - \frac{1}{2}\text{组距}$$

当组距确定后，首先要选定第一组的组中值。一般第一组的组中值以接近于或等于资料中的最小值为好。当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 48.7，所以组中值可取 48.65，因组距为 0.3，因此

第一组的下限应为：$48.65-\frac{0.3}{2}=48.5$；

第一组的上限也就是第二组的下限，应为：$48.65+\frac{3.0}{2}=48.8$；

第二组的上限也就是第三组的下限：$48.8+0.3=49.1$；

……

依次确定各组下限值。为了明确分组界限，各组可只写下限值，后引一波折线的方法表示，如 48.5～，48.8～，49.1～，……。

4. 作次数分布表

分组结束后，可按原始资料顺序，将资料中的每一观测值逐一归组，随后统计每组内所包含的观测值个数，制作次数分布表。一般可将正好等于前一组上限和后一组下限的数据归入后一组。

100 份小包装豆粉净重量的次数分布见表 1－3。

表 1－3　　100 份小包装豆粉净重量的次数分布

组　限	组中值（x）	次数（f）
48.5～	48.65	1
48.8～	48.95	1
49.1～	49.25	6
49.4～	49.55	21
49.7～	49.85	32
50.0～	50.15	23
50.3～	50.45	12
50.6～	50.75	2
50.9～	51.05	1
51.2～	51.35	1

由表 1－3 可以看出，100 份小包装豆粉的净重量多数集中在 49.85g，约占观测值总

个数的1/3，用它来描述小包装豆粉的净重量平均水平，有较强的代表性。净重量小于48.8g及大于51.2g的为极少数。

5. 次数分布图

次数分布用图示的形式表示出来，就是次数分布图。次数分布图主要有直方图、折线图两种。次数分布图以分组数为横坐标，次数为纵坐标绘制。如图1-2和图1-3所示，由次数分布图明显看出100份小包装豆粉的净重量分布情况以及平均净重量。

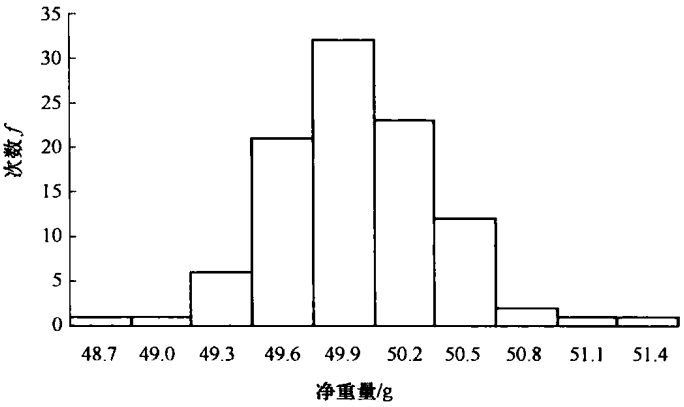


图1-2　100份小包装豆粉的净重量次数分布直方图

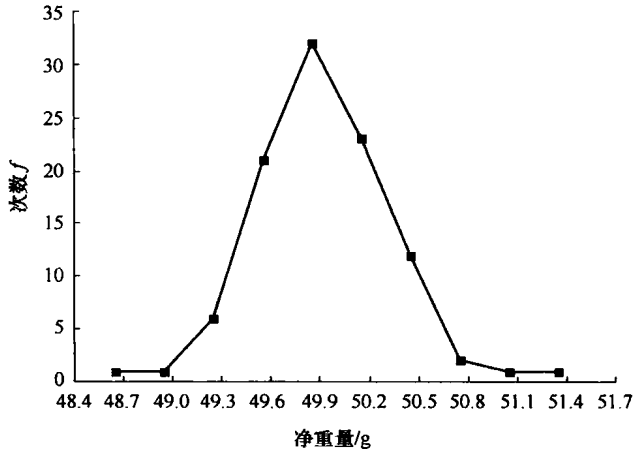


图1-3　100份小包装豆粉的净重量次数分布折线图

第四节　理论分布

一、正态分布

正态分布又称高斯分布，是一种最常见、最重要的连续型随机变量的概率分布。在自然现象中有许多变量取值是服从或近似服从正态分布的，许多统计分析方法都是以正态分布为基础的。此外，还有不少随机变量的概率分布在一定条件下以正态分布为其极限分布。

1. 正态分布的定义

若连续型随机变量 X 的概率密度函数是

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\quad(-\infty<x<+\infty)\tag{1-14}$$

则称随机变量 X 服从平均数为 μ、方差为 σ^2 的正态分布（normal distribution），记作 $X \sim N(\mu, \sigma^2)$。正态分布概率密度曲线如图 1-4 所示。

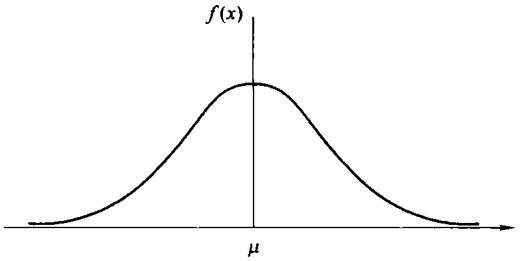


图 1-4　正态分布密度曲线

相应的随机变量 X 概率分布函数为：

$$F(x) = \int_{-\infty}^{x} f(x)\mathrm{d}x = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \quad (-\infty < x < +\infty) \tag{1-15}$$

它反映了随机变量 X 取值落在区间 $(-\infty, x)$ 的概率。

2. 正态分布的性质

（1）正态分布密度曲线是单峰、对称的悬钟形曲线，以 $x=\mu$ 为对称轴。

（2）概率密度函数 $f(x)$ 是非负函数，以 x 轴为渐近线，分布从 $-\infty$ 至 $+\infty$；当 $x \to \pm\infty$，函数 $f(x)$ 曲线接近于 x 轴。

（3）$f(x)$ 在 $x=\mu$ 处有极大值，$f(\mu) = \dfrac{1}{\sigma\sqrt{2\pi}}$。

（4）曲线在 $x=\mu\pm\sigma$ 处各有一个拐点，即曲线在 $(-\infty, \mu-\sigma)$ 和 $(\mu+\sigma, +\infty)$ 区间上是下凸的，在 $[\mu-\sigma, \mu+\sigma]$ 区间内是上凸的。

（5）μ 和 σ^2 是正态分布的两个重要参数，决定着正态分布曲线的位置和形状。

μ 是位置参数，如图 1-5 所示。当 σ 恒定时，μ 愈大，则曲线沿 x 轴愈向右移动；反之，μ 愈小，曲线沿 x 轴愈向左移动。

σ 是形状参数，如图 1-6 所示。当 μ 恒定时，σ 愈大，表示 x 的取值愈离散，曲线愈“胖”；σ 愈小，x 的取值愈集中在 μ 附近，曲线愈“瘦”。

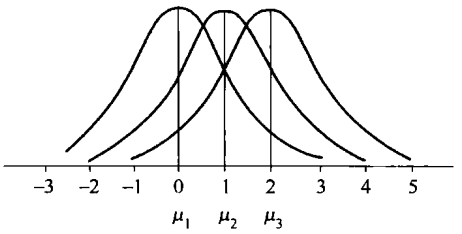


图 1-5　σ 相同而 μ 不同 3 个正态分布的比较

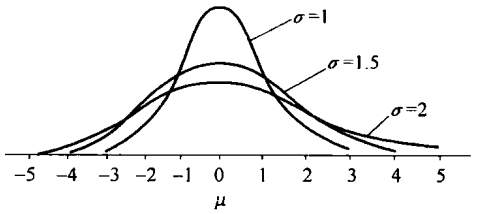


图 1-6　μ 相同而 σ 不同 3 个正态分布的比较

3. 标准正态分布

当正态分布的参数 $\mu=0$，$\sigma^2=1$ 时，称随机变量 X 服从标准正态分布（standard normal distribution），记作 $X \sim N(0, 1)$。其概率密度函数用 $\varphi(x)$ 表示。

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{1-16}$$

相应的分布函数为：

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}} \mathrm{d}x \tag{1-17}$$

标准正态分布密度曲线如图 1－7 所示。

对于任何一个服从正态分布 $N(\mu, \sigma^2)$ 的随机变量 X，都可以通过标准化变换。

$$U=\frac{X-\mu}{\sigma} \qquad (1-18)$$

将其变换为服从标准正态分布的随机变量 U。U 称为标准正态变量或标准正态离差（standard normal deviate）。

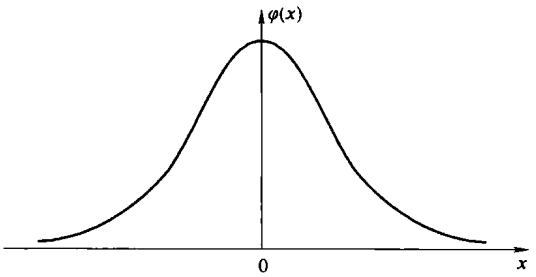


图 1－7　标准正态分布密度曲线

4. 正态分布的概率计算

（1）标准正态分布的概率计算　为了简化标准正态分布函数 $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{x^2}{2}}dx$ 的概率计算，人们编制了标准正态分布函数 $\Phi(x)$ 的数值表，见附表 1。

若 $X\sim N(0, 1)$，对任意 $a<b$ 有

$$\begin{aligned}P(a\leqslant X\leqslant b)&=\int_a^b\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=\int_{-\infty}^b\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx-\int_{-\infty}^a\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx\\&=\Phi(b)-\Phi(a)\end{aligned} \qquad (1-19)$$

[**例 1－2**]　设 $X\sim N(0, 1)$，求①$P(0.5<X<1.5)$；②$P(X<-1.54)$③$P(X>2.50)$

①$P(0.5<X<1.5)=\Phi(1.5)-\Phi(0.5)=0.9332-0.6915=0.2417$

②$P(X<-1.54)=\Phi(-1.54)=1-\Phi(1.54)=1-0.9382=0.0618$

③$P(X>2.50)=1-P(X\leqslant 2.50)=1-\Phi(2.50)=1-0.9938=0.0062$

（2）一般正态分布的概率计算　若随机变量 X 服从正态分布 $N(\mu, \sigma^2)$，则 X 的取值落在任意区间 $[x_1, x_2)$ 的概率，记作 $P(x_1\leqslant X<x_2)$，等于图 1－8 中阴影部分的面积。即：

$$P(x_1\leqslant X<x_2)=\frac{1}{\sigma\sqrt{2\pi}}\int_{x_1}^{x_2}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \qquad (1-20)$$

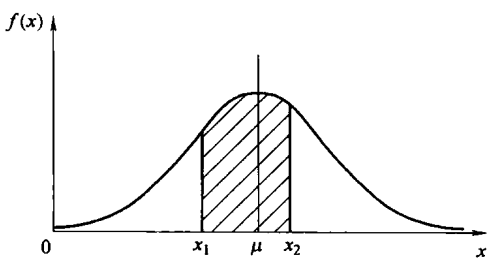


图 1－8　正态分布的概率

对式（1－20）作变换 $u=\frac{x-\mu}{\sigma}$，得 $dx=\sigma du$，故有

$$\begin{aligned}P(x_1\leqslant X<x_2)&=\frac{1}{\sigma\sqrt{2\pi}}\int_{x_1}^{x_2}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{(x_1-\mu)/\sigma}^{(x_2-\mu)/\sigma}e^{-\frac{1}{2}u^2}\sigma du\\&=\frac{1}{\sqrt{2\pi}}\int_{u_1}^{u_2}e^{-\frac{1}{2}u^2}du=\Phi(u_2)-\Phi(u_1)\end{aligned} \qquad (1-21)$$

其中，$u_1=\frac{x_1-\mu}{\sigma}$，$u_2=\frac{x_2-\mu}{\sigma}$

因此，计算一般正态分布的概率时，只要将区间的上下限作适当变换，就可用查标准正态分布的概率表的方法求得概率。