1 问题的重述

1.1 基本情况

我国幅员辽阔，人口众多，每天都消费大量的各种食品。建立完善的食品卫生安全保障体系和风险评估体系关系国计民生和全民身体健康，意义重大。为此，我国曾在 12 个省、自治区、直辖市开展过三次大规模的居民总膳食研究调查，集中力量对众多污染物中少数几种危害面广、后果严重的污染物，如：铅、镉、汞、砷、有机磷、有机氯等实行监控，为开展食品安全评估提供了科学的依据。但是和美国、欧盟等一些发达国家比较，我国食品卫生安全评估体系和风险预警体系还不够完善，存在一些亟待解决的问题。

1.2 要解决的问题

1.2.1 设计抽样调查方案，使得调查方案能够尽量反映全国的实际情况。

1.2.2 建立人群食物摄入量模型。用于估计不同地区、不同性别、不同年龄、不同季节、不同劳动强度、不同经济收入的人群各类食品的一天摄入量。

1.2.3 建立污染物分布模型。用于反映各类食物中各种污染物的分布规律。

1.2.4 建立风险评估模型。用于根据膳食模型和污染物分布模型数据结果，对某一时段食品安全作出评估。

2 问题的假设

2.1 所划分的各类食品包含的品种能够代表当前主流品种。

2.2 根据饮食习惯等划分的全国四区 12 个省能够代表全国的基本情况。

2.3 劳动强度、经济收入分别划分为轻、中、重和低、中、高三种情况。

2.4 食物中含有的主要污染物为：铅、镉、汞、砷、有机磷、有机氯等六种。

2.5 某一地区，某一时间段内，具体某类食物中的食物品种认为其主要污染物含量相同。

2.6 例行检测抽调的样本数量远远大于偶然性检测抽调的样本数量。

2.7 文中调查、检索、引用的数据科学合理并具有一定的代表性。

3 符号说明

A_i :划分的第 i 个大区；

B_i ：抽调分析的第 i 个省；

b_i ：第 B_i 个省抽调分析的 b_i 个单位；

C_i ：省抽调分析的第 i 个下属单位；

c_i ：第 C_i 个下属单位抽调分析的 c_i 个再下属单位；

C_T ：总费用固定、估计量的方差最少时的费用函数；

γ_h：第h层样本的抽样比；

P：各下属单位抽样得到的总体样本概率；

P_i：按照行政区域分层得到的第i层抽样概率；

M_i：抽调的第i个样本；

m：从各大区抽取各省（自治区）的样本个数；

n：从各省（自治区）抽取各下属单位的样本个数；

N_1, N, N_2：神经网络的输入单元数层数、隐含层单元数和输出层单元数；

WR_i：第i类污染物；（铅、镉、有机磷、有机氯、汞、砷）；

W_i：第i层样本的层权值，也用W_h表示；

S_h：第i层样本的层方差；

$f\left(W^1\right)$：误差代价函数；

$L(a_0, a_1, a_2 \cdots\cdots a_n)$：多项回归分析的似然函数；

P_q：表示北一区、春季、谷类食品关于铅元素的摄入量的概率值；

P_s：选中W^1为下一代个体的次数；

Y：样本总体总和；

$\hat{Y}_{PPS}$：总体总和Y的估计量；

ZH_i：第i类污染物国家安全标准；

YFD_{ij}：第i类食物第j类污染物99.999%的右分位点；

$f_i(x)$：第i类食品污染物含量分布密度函数；

(a,b)：分布验证检测区间；

δ：分布验证相似度临界值；

F（x）：抽样样本总体分布函数；

4 抽样调查方案的设计

4．1 抽样调查方案设计分析

居民膳食摄入量调查设计面积广，用户众多，不可能机械的进行普遍地毯式调查，即使调查的样本很多，上百万甚至上千万，但对于13亿中国人口来说仍然是沧海一粟。

在这种情况下，如何设计抽样调查方案，使得调查的少样本数据具有普遍的代表性显得尤为重要。同时，食物摄入量模型的建立不能无中生有，必须依据调查的相关数据建立，因此，只有样本数据具有代表性且符合实际情况，才有可能建立具有全国通用型的模型，较准确的估计不同情况下人群各类食品的一天摄入量。

抽样调查方案有以下特点：一是总体在不断变化，这给抽样调查带来两个方面的困难：一是抽样框难于确定不易获得有关抽样样本；二是由于小范围内对总量估计的高估和低估。二是需要满足多层次的推断。抽样调查不只是满足总体这个层次的估计，而且也希望尽量满足多个层次的估计需要。

抽样调查方案的设计应充分考虑不同地区、不同性别、不同年龄、不同季节、不同劳动强度、不同经济收入几种情况下人群不同种类食物的摄入量。为了尽量减少抽样调查的工作量，可以根据一定的标准、原则和常识对以上六个变化的量进行分层分析，使每一层在一定范围内具有较高的代表性。因此，我们需要设计一种多变量多层次抽样调查方案，使得这样的抽样调查方案既考虑了样本误差的控制，又能显著降低样本调查的样本容量。它最大的特点是可以将大样本问题转换成相对小的样本问题，减少了工作量和对样本数据绝对数量的要求。

大样本向小样本转换，为了保证小样本能够代表大样本说明问题，下一步需要对样本数据进行目标量及方差估计，如果样本符合标准，则相信根据设计的抽样调查方案抽得的样本数据具有典型性和代表性；如果不符合则采用分层不放回追加，虽然相应的估计方法较为复杂，但这种追加方法所需追加样本较少即可使得样本转换满足标准。

4．2 多变量多层次食物摄入量混合抽样调查方案设计

4.2.1 抽样调查方案设计的目标

抽样调查是一项技术，更是一项艺术，创新性和应用性很强。多变量多层次混合抽样调查方案设计的目标是期望用尽可能少的样本高精度地估计总体的数量特征。多层次均匀试验设计所选择的抽调点在实验范围内分布均匀，且抽调样本数目大大地小于其他实验方法的抽调样本数目。

4.2.2 基本思路

分层实验设计关键是如何进行分层，将具有相同或者相似属性的对象归为一类，视为同一层次。为此，根据中国总膳食研究及相关文献资料对食物进行分层，然后确定目标变量，显然，本抽样是多变量多层次的混合抽样调查，方案设计应兼顾考虑，不可偏废。样本抽查要有比例，例如根据全国年龄分布比例，劳动强度分布比例、经济收入分布比例等确定各种类型的群体抽样比例。因此，第一步确定分层标志的选择和样本抽查比例；第二步是采用混合多层次多变量抽样；第三步是分层随机抽样；最后进行抽样误差估计。

4.2.3 抽样调查方案的设计

4.2.3.1 混合分层设计

（1）食物分层

根据中国营养学会提供的居民平衡膳食宝塔(图 4-1)和美国 2005 年膳食金字塔(图 4-2），将食品进行分层如表 4-1：

图 4-1 中国居民平衡膳食宝塔

图 4-2 美国 2005 年膳食金字塔

表 4-1 食物分层表及各层包含食品

编号 (A)	食品分层	包　　括
A1	谷　　类	面粉、大米、玉米粉、小麦、高粱等
A2	蔬　　菜	根菜类（萝卜等）、鲜豆类（豆角、蚕豆等）茄果瓜菜类（茄子、黄瓜等）、葱蒜类、根、茎、叶类（白菜、芹菜等）等
A3	水　　果	苹果、梨、桔子等
A4	肉　　类	猪、牛、羊、鸡肉等及其制品
A5	蛋　　类	鸡蛋、鸭蛋、鹌鹑蛋及其制品等
A6	水　　产	鱼、虾、蟹、贝克类等及其制品
A7	乳　　类	奶粉、酸奶、奶酪酥油等及其制品
A8	薯　　类	土豆、山芋、洋芋等及其制品
A9	豆　　类	黄豆、豆腐、豆制品等
A10	饮 料 类	碳酸饮料、茶饮料等
A11	酒 饮 料	发酵酒（啤酒、黄酒等）、蒸馏酒（曲酒、二锅头等）、露酒（密酒、香雪酒等）

（2）地区分层

我国地域辽阔，各地居民饮食习惯差异较大，膳食结构不尽相同，因此，分地区对居民各类食品摄入量进行研究非常必要。经过大量检索文献和分析资料的基础上我们发现：中国南方、北方两大地区饮食各有特点而又存在较大的差异。例如北方主食、水果、奶制品以及油脂的摄入量高于南方；南方主食品种比较单调，以大米为主，但豆类、蔬菜、肉类、水产品和酒类摄入量明显高于北方。因此，我们将地区进行如下划分：

表 4-2 地区分层及试点

南北地区	划分地区	试点地区
北方	北一	黑龙江、辽宁、河北
	北二	河南、陕西、宁夏
南方	南一	上海、福建、广东
	南二	湖北、四川、广西

（3）经济收入分层

经济收入水平划分有多种标准，这里我们以家庭人均收入水平作为确定收入者的唯一标准。这种方法指向准确，方法简便，概括性强，可以排除各类复杂因素的干扰，不会引起较多争议。收入是综合性的，是基本前提，人们能够获得某种程度的收入，往往决定于他们从事的是什么职业，达到何种教育程度和技能水平，等等。同时，一定的收入水平在某种程度上又可以决定人的生活方式、受教育程度、积累财产及各类生产要素的能力，甚至思想观念的倾向，等等。采用城乡居民家庭人均收入标准，还可以真实反映每个居民或每个家庭成员生活与发展的实际条件，且符合国际通行做法。

在对经济收入水平进行划分时，我们只要得出中等收入者的标准，即可推出高收入者、低收入者的标准。在我国，现阶段由于城乡发展差距和收入差距巨大，因而如何确定中等收入者的标准，是一个难度很大的问题。关于中等收入者标准的论述很多，也有一定的道理，但至今还没有一个统一的标准。当然，无论如何，在不同发展阶段，中等收入者比重不应是固定的，而应是不同的。在模型中，我们将中等收入者的标准定在10000 元以上（包含 10000 元）、40000 元以下（不包含 40000）。原因有二：一是中等收入者收入的起点标准应比目前我国人均国内生产总值略高一些，接近职工年平均工资水平，能保证人均食品消费支出比例低于 25%，而且要适当贴近国际平均的中等收入标准线。这样的家庭人均收入水平才能基本支持人力资源发展或人的素质较全面提高。2002 年，我国人均国内生产总值约 9000 元，职工年平均工资约 12000 元，能够支持 25%左右食品消费的城镇家庭人均收入水平也大致在 10000 元左右。此外，近期世界银行的一项研究表明，全球中等收入层的人均收入起点标准为 3470 美元，经购买力平价调整后大约相当于 14500 元人民币。把我国中等收入者家庭人均收入起点标准定在其三分之二左右的水平上比较合适。我国城乡居民家庭收入调查中中低收入与最高收入家庭人均年收入的实际差别在 4 倍左右；我国高收入家庭人均年收入标准应能大致达到发达国家人均 GDP 的初始标准，世界银行近期的研究结果将全球中等收入层的高限确定为 8000 美元，通过购买力平价调整，应在 40000 元人民币左右；将 40000 元作为高收入者的起点标准，其人均年收入可超过 10 万元，接近目前多数人对高收入者的认定标准。

依据上述标准，我们把经济收入水平划分为高、中、低三个层次。具体划分如下：

表 4-3 经济收入分层及分层便准

经济收入水平	收入（元）
高	＜10000
中	10000～40000
低	≥40000

（4）劳动强度分层

GB 3869-1997《体力劳动强度分级》标准将体力劳动强度分为Ⅰ、Ⅱ、Ⅲ、Ⅳ级。该标准适用于以体力活动为主的劳动，但在模型建立过程中，我们还需要考虑脑力劳动与食物摄入量的关系。因此，我们依据《成人活动水平（PAL）分级》，综合考虑体力劳动和脑力劳动，将劳动强度分为轻、中、重三个级别。如下表所示：

表 4-4 劳动强度分层及分层标准

劳动强度	职业工作及时间分配	工作内容举例
轻	75%时间坐或站立 25%时间站着活动	办公室工作、修理电器钟表、售货员 酒店服务员、化学实验操作、讲课等
中	25%时间坐或站立	学生日常活动、机动车驾驶、电工安装、

	75%时间特殊职业活动	车床操作、金工切割等
重	40%时间坐或站立 60%时间特殊职业活动	非机械化农业劳动、炼钢、舞蹈、 体育运动、装卸、采矿等

4.2.3.2多变量多层混合抽样调查方案

（1）方案论述

为使抽样数据可以同时满足国家统计需要，实现小样本数据对总体分布中的一些未知因素做出推断的目的，抽样方法必须满足以下条件：

1）代表性。总体中的每一个体都有同样机会被抽入样本，这意味着样本中每个个体与所考察的总体具有相同的分布，因此，任意一个样本的个体都具有代表性；

2）独立性。样本中每个个体取什么值并不影响其它个体取什么值。这就是说，样本中各个体之间是相互独立的随机变量。

由简单随机抽样所得到的样本称为简单随机样本，它可以用与总体同分布的n个相互独立的随机变量表示。假设总体的分布函数为F（x），则其简单随机样本的联合分布函数即为：

$$F（x_1，x_2，……x_n)=F（x_1）F（x_2）……F（x_n）\qquad （公式\ 4\text{-}1）$$

当总体数量较大或所抽样本数量在总体所占比例较小时，不放回抽样就可获得较理想的简单随机样本。当样本容量较小时，只有进行放回抽样才能获得简单随机样本，但在实际抽样中，获得简单随机样本不难，难的是人为干扰因素的影响，如不同年龄涉及到与经济收入矛盾的问题等等。本题要获取全国食品卫生安全保障体系大样本总体特征及其参数的信息，由于此类项目在每个样本上数据调查采样量比较大，预算的范围决定了最后接受调查的样本容量相对很大，有限的预算和时间限制决定了我们选择采用普查和抽样的方式——“多变量多层次混合抽样”获得抽样调查数据。

（2）方法论述

“多变量多层次混合抽样”调查方法是在结合考虑了预算和精确度后进行的，它的混合分层包括：首先，采用混合多级抽样，获取一个比较大的样本，针对不同地区、不同性别、不同年龄、不同季节、不同劳动强度、不同经济收入的人群食物摄入量数据分布，得到一些总体的辅助信息，如根据总膳食研究的分区原则和方法，将全国地区大致分为南一、南二、北一、北二4个大区，人口覆盖率达到47％的4个大区分别为：北方一区(黑龙江省、辽宁省、河北省)；北方二区(河南省、宁夏回族自治区、陕西省)；南方一区(江西省、上海市、福建省)；南方二区(湖北省、四川省、广西壮族自治区)，其中选点的总原则是要使所选的点能代表本省人民的饮食习惯、营养状况和实际膳食结构；其次，根据各典型地区主要调查对象数据，得到大样本的相关信息（性别比例、收入情况等等)，根据相关信息对大样本进行分层，在大样本中每一层随机产生每一层小样本，力求解决预算和精度上达到要求的目的。

第一步：混合多级抽样

我国当前的行政区域管理体制具有很强的地域性，即：全国——各大区——各省（自治区）——各下属单位。大样本的确定可以根据四级抽样进行，而在每一级采取PPS(有放回的按与单元“大小”成比例的概率来抽取样本的方法)抽样法。具体操作如下：

设总样本共N人，A个大区，B个省（自治区），C个下属单位,欲从总样本中最后得到n个样本进行主要调查。

1）为了保证样本的代表性，一级抽样(从全国抽取各大区)和二级抽样(从各大区中抽取各省（自治区）)采取抽取所有单元的方法。

2）三级抽样(从各省（自治区）抽取各下属单位)采用PPS法抽样，抽取指定数目的各下属单位，下属单位数目的确定方法由混合多级抽样的样本量和经费确定。

在制作抽样框时，访问员登记的抽样框的每个抽样单元可视做随机顺序排列的。为使抽样具有随机性，在现实中通常采用等距抽样，使其接近随机抽样。从各大区中抽取各省（自治区）时可采用循环等距抽样法。等距抽样的随机起点通常在第一个k值中利用随机数表产生，其中k为抽样间距。

如果设各省（自治区）有B个，共R个下属单位；每个省（自治区）用B_i表示，每个B_i中有b_i个单位；C个各下属单位中属于B_i的各下属单位用C_i表示（i=1，2，3，……），每个C中有c_i个单位。为了使抽样接近随机抽样，在抽样时必须保证抽样样本中的每个下属单位被抽中的概率相等，即每个各下属单位被抽中的概率为：$P = n/R$。设抽中各大区的概率P_1，抽中各省（自治区）的概率为P_2，各省（自治区）中抽取各下属单位的概率为P_3，使得$P = P_1P_2P_3$。

①如果要从各大区中抽取各省（自治区）m个，则每个各大区抽取各省（自治区）个数为：$m \times b_i / R$，确定每个各大区中抽取各省（自治区）个数后，再将各大区中的各省（自治区）编号排序，采用PPS法用等距抽样抽出各省（自治区）。

②如果从选中的各省（自治区）中抽取n个各下属单位，则每个各省（自治区）抽取各下属单位个数为$\frac{nR}{mb_1}$，确定每个各省（自治区）中抽取的各下属单位个数后，每个各省（自治区）的各下属单位编号，采用等距抽样抽取各下属单位，如下表：

表 4-5混合多级型抽样第一阶段

第一阶段（从全国中抽取各大区，抽取全部）				
按行政区域分层	A_1大区	A_2大区	A_3大区	A_4大区
各层应抽各大区数	全部	全部	全部	全部
各大区被抽中的概率	1	1	1	1

表 4-6混合多级型抽样第一阶段

第二阶段（从各大区中抽取各省（自治区），设抽取m个）				
按行政区域分层	B_1各省（自治区）	B_2各省（自治区）	B_3各省（自治区）	…
各层应抽各省（自治区）数	$m \times b_1 / R$	$m \times b_2 / R$	$m \times b_3 / R$	
各省（自治区）被抽中的概率	$C_1 \times m \times b_1 / R^2$	$C_2 \times m \times b_2 / R^2$	$C_3 \times m \times b_3 / R^2$	

表 4-7混合多级型抽样第三阶段

第三阶段（从各省（自治区）中抽取各下属单位，设抽取n个，在第二阶段被抽中的各省（自治区）				
按行政区域分层	C_1 各省（自治区）	C_2 各省（自治区）	C_3 各省（自治区）	…
各层应抽各下属单位数	$\frac{nR}{mb_1}$	$\frac{nR}{mb_2}$	$\frac{nR}{mb_3}$	
各下属单位被抽中的概率	$\frac{nR}{C_1mb_1}$	$\frac{nR}{C_1mb_1}$	$\frac{nR}{C_1mb_1}$	

第二步：分层随机抽样

在我们对第一步样本进行调查后，得到大样本的相关信息(比如性别比例、年龄结构、收入结构、受教育情况等)。根据相关信息对大样本进行分层，在大样本中每一层随机产生每一层小样本，如下表所示。

表 4-8 大小样本示意表

大样本	大样本M			
分层后的大样本	M_1	M_2	M_3	……
小样本	m_1	m_2	m_3	……

其中：

$m_1 = m * W_1$； $m_2 = m * W_2$； $m_3 = m * W_3$； …… $m = m_1 + m_2 + m_3 + \cdots\cdots$

其中，W_i 为层权，通常由大样本和小样本的比例决定，在实际调查中通常要考虑许多实际因素，如层内方差、各下属单位规模等，所以实际上各层权不尽相同。为使抽样接近随机，在分层后的大样本中抽取小样本时，采用等距抽样保证其随机性，由于两次样本的抽样方法都接近随机抽样或高于随机抽样，故最后的小样本可看作随机样本，便于计算机操作。

第三步：分层标志的选择

第二步样本从第一步样本中分层后产生，为提高调查精度，通常采用最优分配原则，在确定分层标志后，层权由分层比例和层内方差决定．为使分层抽样间方差尽量大，选择好分层标志十分关键。根据调查的最终目的，即个人收视行为的调查，可选定家庭收入作为分层标志。在具体分层中，采用等分分层变量(标志)分布的累计平方根的最优分层方法，即简称累积平方根法。假设以收入Y作为分层标志，对于最优(纳曼)分配情形，目标是使估计量的方差极小化，也就是使 $\sum_h W_h S_h$ （W_h 为层权重，S_h 为层方差)最小化。

设Y的分布频率为 $f(y)$，在给定的层中，将 $f(y)$ 看做常数，也就是说服从均匀分布，设层的分点为 $y_1, y_2, \cdots\cdots, y_{L-1}$，则

$$\sum_h W_h S_h \approx \frac{1}{\sqrt{12}} \sum_{h=1}^{L} f_h \left(y_h - y_{h-1}\right)^2 \approx \frac{1}{\sqrt{12}} \sum_{h=1}^{L} \left(Z_h - Z_{h-1}\right)^2 \qquad \text{（公式 4-2）}$$

其中，

$$Z_h = \int_{y_t}^{y_h} \sqrt{f(t)}dt \qquad \text{（公式 4-3）}$$

这里 y_0 是第一层的起点，可以证明当 $Z_h - Z_{h-1}$ 都相等时，式（1）的值达到最小，因此只要 $f(y)$ 已知，就可按 $\sqrt{f(y)}$ 的累积值来确定最优分层的分点。

（3）抽样调查比例的确定

1）各区人口数量及比例

地　区	出生率（‰）	死亡率（‰）	自增率（‰）	年底人口数（万人）
全　国	12.09	6.81	5.28	131448
北　京	6.26	4.97	1.29	1581
天　津	7.67	6.07	1.60	1075
河　北	12.82	6.59	6.23	6898
山　西	11.48	5.73	5.75	3375
内蒙古	9.87	5.91	3.96	2397
辽　宁	6.40	5.30	1.10	4271
吉　林	7.67	5.00	2.67	2723
黑龙江	7.57	5.18	2.39	3823
上　海	7.47	5.89	1.58	1815
江　苏	9.36	7.08	2.28	7550
浙　江	10.29	5.42	4.87	4980
安　徽	12.60	6.30	6.30	6110
福　建	12.00	5.75	6.25	3558
江　西	13.80	6.01	7.79	4339
山　东	11.60	6.10	5.50	9309
河　南	11.59	6.27	5.32	9392
湖　北	9.08	5.95	3.13	5693
湖　南	11.92	6.73	5.19	6342
广　东	11.78	4.49	7.29	9304
广　西	14.44	6.10	8.34	4719
海　南	14.59	5.73	8.86	836
重　庆	9.90	6.50	3.40	2808
四　川	9.14	6.28	2.86	8169
贵　州	13.97	6.71	7.26	3757
云　南	13.20	6.30	6.90	4483
西　藏	17.40	5.70	11.70	281
陕　西	10.19	6.15	4.04	3735
甘　肃	12.86	6.62	6.24	2606
青　海	15.24	6.27	8.97	548
宁　夏	15.53	4.84	10.69	604
新　疆	15.79	5.03	10.76	2050

图 4-3 2006 年国家统计局各地区人口变动情况统计结果

表 4-9 各省人口数量及所占比例

地区划分	主要各省	人口数量（万人）	各省人口所占比例
北一区	黑龙江	3823	6.17%
	辽宁	4271	6.89%
	河北	6898	11.13%
北二区	河南	9392	15.15%
	陕西	3735	6.03%
	宁夏	604	0.97%
南一区	上海	1815	2.93%

	福建	3558	5.74%
	广东	9304	15.01%
南二区	湖北	5693	9.19%
	四川	8169	13.18%
	广西	4719	7.61%

各省人口分布图如下：

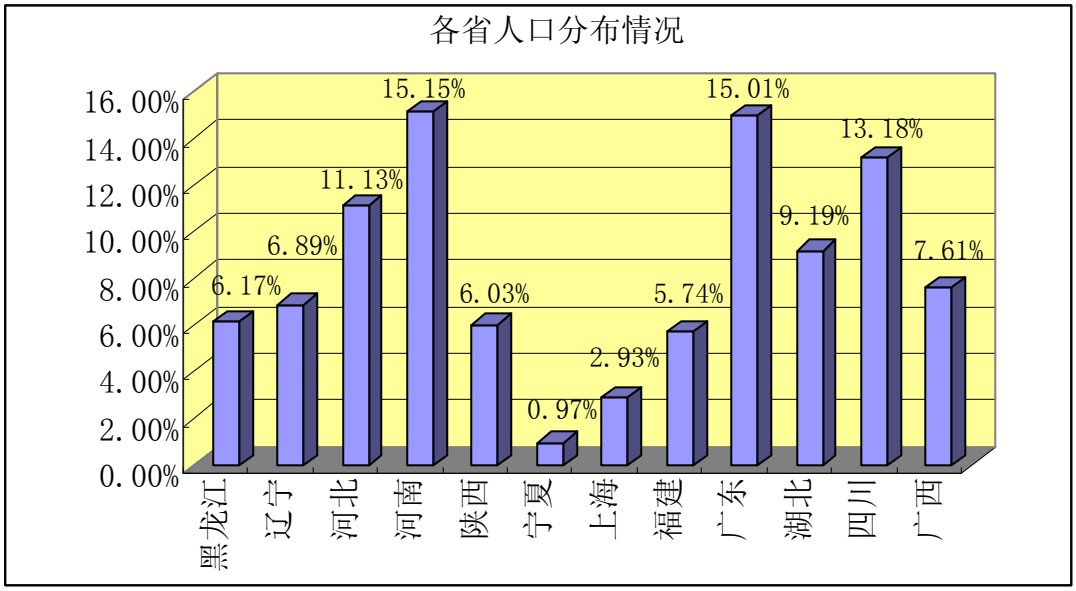


图4-4 各省人口分布柱状图

2）年龄段统计结果及比例

表 4-10 年龄段统计

年龄阶段	年龄（岁）	所占比例（%）
少年	≤14	23.85%
青年	15-44	49.35%
中年	45-59	15.74%
老年	≥60	11.06%

各年龄段人口分布图如下所示：

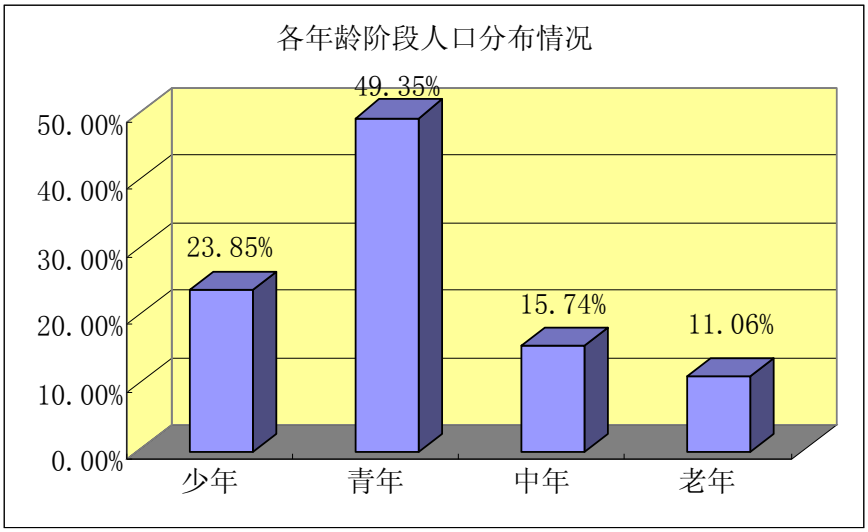


图4-5 各年龄段人口分布比例

3）男女比例统计结果及比例

表 4-11男女比例

性别	数量（万人）	比例（%）
男	67309	51.53
女	63319	48.47

4）经济收入统计结果及比例

据全国总工会 2005 年对 20 个市（区）1000 个各种所有制企业以及 1 万名职工的问卷调查的结果，并查阅了大量资料，我们得出高、中、低水平收入者所占比重如下表所示：

表 4-12 经济收入比例

经济收入水平	比重（%）
高	13
中	58
低	29

5）劳动强度统计结果及比例

据不完全统计，劳动强度为轻、中、重级的劳动者所占比重如下表所示：

表 4-13 劳动强度比例

劳动强度	比重（%）
轻	21
中	62
重	17

（4）抽样误差估计和样本量的确定

第一步，样本抽样误差的估计

在市场研究中，为了便于计算，多级混合抽样产生的第一步样本的抽样误差的估计可视为以下抽样误差的估计：

1)初级单元(从各大区抽取各省（自治区）)看作PPS抽样，即与单位规模成比例抽样；

2)次级单元(从各省（自治区）抽取各下属单位)看作是简单随机抽样。

设Y_{ij}表示总体中第i个初级单元中第j个次级单元的指标值，

$$i=1,2,\cdots,N;j=1,2,\cdots,M_i$$

又$M^0=\sum_{i=1}^{N}M_i$是总体中次级单元的总数，y_{ij}表示样本中第i个初级单元中第j个次级单元的观测值，$i=1,2,\cdots n;j=1,2,\cdots,m_i$，其中n与$m_i$分别是第一阶抽样和第二阶抽样的样本量，而$f_1=\frac{n}{N}$，$f_{2i}=\frac{m_i}{M_i}$分别是抽样比，另外记总体及样本各级总和、均值与方差如下：

$$Y_i=\sum_{j=1}^{M_i}Y_{ij} \quad （公式 4-4）$$

$$y_i=\sum_{j=1}^{m_i}y_{ij} \quad （公式 4-5）$$

$$Y=\sum_{i=1}^{N}Y_{ij} \quad （公式 4-6）$$

$$y=\sum_{i=1}^{n}y_{ij} \quad （公式 4-7）$$

$$\overline{Y_i}=Y_i/M^i \quad （公式 4-8）$$

$$\overline{y_i}=y_i/m_i \quad （公式 4-9）$$

$$S_{2i}^2=\frac{1}{M_i-1}\sum_{j=1}^{M_i}\left(Y_{ij}-\overline{Y_i}\right)^2 \quad （公式 4-10）$$

$$s_{2i}^2=\frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(y_{ij}-y_i\right)^2 \quad （公式 4-11）$$

$$\overline{\overline{Y}}=Y/M^0 \quad （公式 4-12）$$

$$\overline{\overline{Y}}=y/\sum_{i=1}^{n}m_i \quad （公式 4-13）$$

$$\overline{Y}=Y/N \quad （公式 4-14）$$

$$y=y/n \quad （公式 4-15）$$

对于初级单元的PPS抽样，因次级单元是简单随机的，则此总体总和Y的估计量即为：

$$\hat{Y}_{PPS}=\frac{M_0}{n}\sum_{i=1}^{n}\overline{y_i} \quad （公式 4-16）$$

第二步，样本抽样误差的估计

记第一步样本为n，n_h是属于第一步样本中h层的单元数。

记第二步样本为n'，n_h'是属于第二步样本中h层的单元数，记L为层数。则：$w_h'=\frac{n_h}{n}$是总体中实际层权的一个无偏估计。

为使估计量的方差最少，层权可以根据与方差和比例乘积成正比（最优分配）W_kS_k的方法取得，即

$$w_h = \frac{\sum_h w_h' S_h}{\sum_h S_h} \qquad \text{（公式 4-17）}$$

记Y^{hj}是第二步样本h层第j单元的观测值，$j = 1,2,3,\cdots\cdots n_h'; h = 1,2,3,\cdots L$

则h层的平均数

$$\bar{y}_{\mathrm{h}}' = \frac{1}{n_k^1}\sum_{j=1}^{n_k} y_{hj} \qquad \text{（公式 4-18）}$$

是第一步样本h层平均数的无偏估计，故总体均值$\bar{Y}$的二重分层抽样的估计量为：

$$\overline{y_{std}} = \sum_{h=1}^{L} w_h \overline{y_h'} \qquad \text{（公式 4-19）}$$

其方差的一个近似无偏估计为：

$$v\left(\overline{y_{std}}\right) = \sum_n \left(\frac{1}{n_h} - \frac{1}{N}\right) w_h^2 s_h^2 + \left(\frac{1}{n} - \frac{1}{N}\right) \sum_h w_h \left(\overline{y_h'} - \overline{y_{std}}\right)^2 \qquad \text{（公式 4-20）}$$

其中，s_h^2是第二步样本的方差。

第三步，二重样本的最优分配

在二重抽样中有两次抽样，每次抽样都要花费一定费用，第一步抽样的样本量n越大，则对辅助信息的了解和估计就越精确，对改善最终估计量，减少方差越有帮助，然而，另一方面，第一步抽样占用过多的费用必然影响第二步样本n'的抽样，使其不能抽太多的样本，从而还是会影响估计量的精度，增大方差。由此存在两次样本量的最优分配问题。在总费用固定时，估计量的方差最少，定义费用函数为：

$$C_T = c_1 n + \sum c_{2h} n_h' \qquad \text{（公式 4-21）}$$

其中n_h'是属于h层的样本单元数，c_1是第一步样本平均每个单元的调查费用，c_{2h}是第二步抽样中h层的平均每个单元的调查费用，设定每层抽样比为γ_h，期望费用

$$C_T' = c_1 n + n\sum c_{2h}\gamma_h W_h \qquad \text{（公式 4-22）}$$

和抽样比：

$$\gamma_h = S_h \sqrt{\frac{c_1}{c_{2h}\left(S^2 - \sum_h W_h S_h^2\right)}} \qquad \text{（公式 4-23）}$$

估计S^2，S_h^2和W_h值（或采用经验数据），可计算抽样比，再通过费用确定样本量。本文提供的解决样本量确定问题的方法——多变量多层混合抽样调查方案，可能存在在

第一步样本中的愿意合作进行基础研究调查，而在选取第二层样本进行深入调查时候，调查对象可能因为某些原因不愿意继续合作，则缺少的样本可以随机从总体中产生，在调查后采用事后分层(抽样后分层)的方法，即从总体中采用多级混合抽样随机产生，产生后再按照某个指标(通过第一步样本确定的指标)的分层原则对其进行分层，然后再利用事后分层的方法计算其抽样误差。考虑到样本的疲劳和搬迁等因素，实际中的少数样本可在轮换中不断调整和修正，使其满足辅助调研预测得到的总体的结构。

5 基于 BP 神经网络的食物摄入量模型

神经网络是由大量神经元相互作用形成的网络，是一个高度非线性动力学系统，它通过简单的神经元的连接，反映出来的神经网络的动态行为却是十分复杂的，可以表达出实际物理世界的各种现象，是抽象、简化与模拟的人工信息处理模型。神经网络系统具有学习能力、自适应能力、自组织能力、容错与自修复能力、知识表示能力及模式存储、检索能力，它的特点主要是下列几点：

1)并行分布处理；

2)高度鲁棒性和容错能力；

3)分布存储及学习能力；

4)能充分逼近复杂的非线性关系。

人工神经网络的模型现在有数十种之多，包括 BP 网络、Hopfield 网络、ART 网络和 Kohonen 网络。其中以在 1986 年 Rumelhart 等提出的误差反向传播法，即 BP(Error Back Propagation)法影响最为广泛。直到今天，BP 算法是目前最广泛应用的神经网络学习算法之一，常应用于系统辨识、模式识别、智能控制等领域，在自动控制中是最有用的学习算法。这种算法可以对网络中各层的权系数进行修正，故适用于多层网络的学习。

5.1 建立 BP 神经网络模型的主要步骤

5.1.1 BP 神经网络的基本结构

BP 神经网络由输入层、隐含层、输出层构成，基本结构示意图如下图所示：

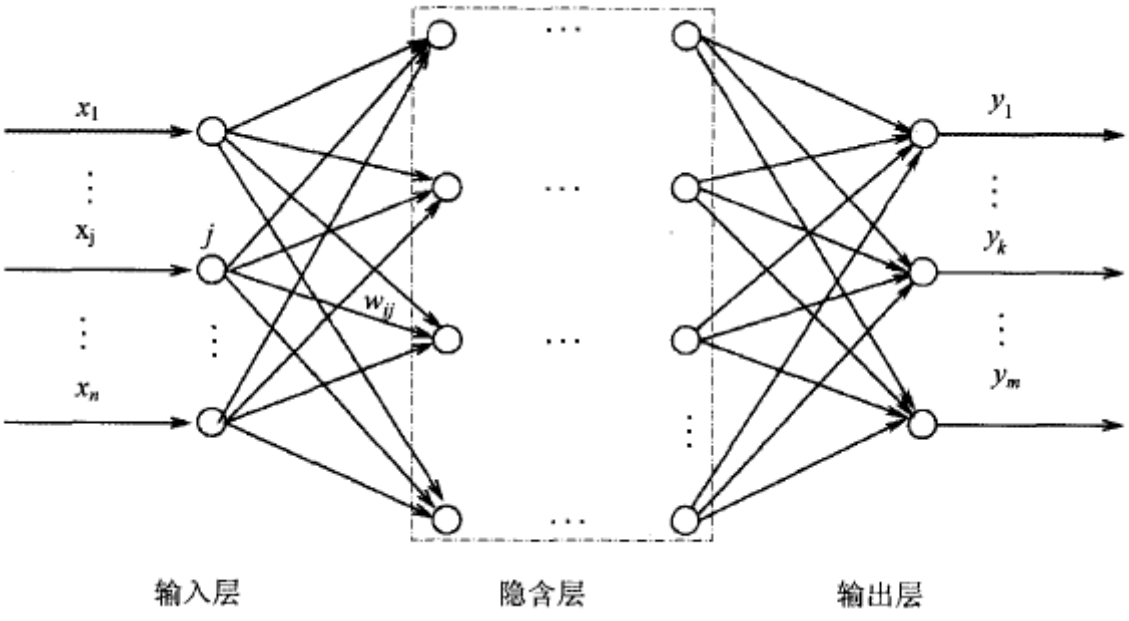


图 5-1　BP 网络拓扑结构示意图

确定 BP 神经网络的结构也就是确定输入层的单元数、隐层数和隐层单元数、输出

层单元数。输入层、输出层的单元数由输入输出数据项决定，而隐层数和隐层单元数的确定却没有比较固定的方法，一般由经验而定。

隐层神经元个数的确定要慎重，神经元个数过多，会使网络变得非常敏感：输入项发生轻微的变化会导致输出结果发生很大幅度的变化；神经元个数过少，会使网络的学习能力降低，甚至有时无法训练网络达到满意的精度。一般需要多次的反复修改试验才能确定最佳的神经元个数。本文参考经验公式 $N=\log_2(N_1+N_2)/2$，来确定初始网络结构中隐层神经元的个数，然后在仿真试验中逐步调整，从而确定最优隐层神经元个数。其中：N_1 为输入神经元个数，N_2 为输出神经元个数，N 为隐层神经元个数。

5.1.2 学习速度、初始权值的选取

较大的学习速度可以提高训练速度，缩短训练时间，但可能会错过全局极小点；较小的学习速度会大大增加网络的训练时间，所以学习速度的设置应该适中。一般情况下如果误差平方和 $\sum e^2$ 下降很快，则说明学习速度合适，如果误差平方和 $\sum e^2$ 出现震荡，则说明学习速度过大，应该调小学习速度。对于较为复杂的网络，误差曲面也会很复杂，在有些部位误差曲面变化很平缓，需要较大的学习速度，而在有些部位，误差曲面变化急剧，需要较小的学习速度，故一般对于复杂网络采取可变的学习速度，在误差曲面变化平缓区域自动增大学习速度，在误差曲面急剧变化的区域自动减小学习速度，实现算法为：

$$W(t+1)=W(t)+\alpha(t)D(t) \qquad \text{（公式 5-1）}$$

其中：

$$\alpha(t)=2^{\lambda}\alpha(t-1) \qquad \text{（公式 5-2）}$$

$$\lambda=sign\left[D(t)D(t-1)\right] \qquad \text{（公式 5-3）}$$

当连续两次迭代其梯度方向相同时，表明下降太慢，这时可使步长加倍；当连续两次迭代其梯度方向相反时，表明下降过快,这时可使步长减半。

初始权值选取的优劣直接关系到网络的训练质量，选取不合理可能会使网络收敛到局部极小点。初始权值的选取一般情况下应该注意两点：第一，初始权值大小要适中，一般情况下应选为均匀分布的小数经验值，约为 (-2.4/F,2.4/F) 之间，其中 F 为所连单元的输入层节点数；第二，初始权值尽量避免全部设置相同，这样容易使网络陷入局部极小点[2]。

5.1.3 训练 BP 神经网络

借助 Matlab 仿真软件对 BP 神经网络进行训练，整个训练的过程也就是不断调整神经元权值的过程，训练次数的确定以及停止训练的时间取决于误差 $\sum e^2$ 是否达到满意的值。如果无论怎么训练也不能使 $\sum e^2$ 达到满意的要求，则需要重新调整网络的结构，重新进行训练，看 $\sum e^2$ 能否达到满意的要求，如果仍不能满足要求，则继续调整网络结构重新训练，直到误差达到满意的要求。

5.1.4 检验 BP 神经网络的训练结果

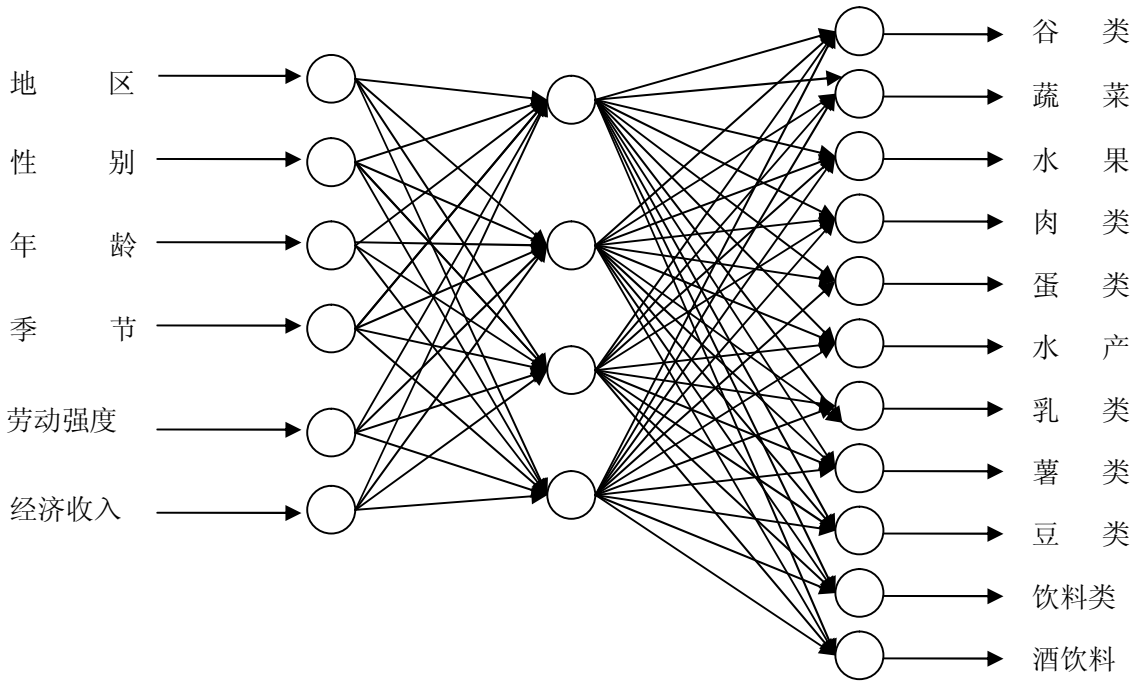
用测试样本检验训练后的 BP 神经网络，将网络的输出和测试样本的期望输出进行比较。如果误差在允许的范围内，则接受训练后的网络，如果误差超过了允许的范围，则重新选取样本对网络进行训练

5.2 建立 BP 神经网络人群食物摄入量模型并仿真

5.2.1 建立 BP 神经网络人群食物摄入量模型

人群食物摄入量模型用于估计不同地区、不同性别、不同年龄、不同季节、不同劳动强度、不同经济收入几种情况下人群不同种类食物的摄入量。因此对于本模型来说，神经网络的输入层又 6 个单元，分别为：地区、性别、年龄、季节、劳动强度、经济收入。输出层单元根据中国总膳食结构和设计的多变量多层次混合抽样调查方案中食物分层确定，因此，输出层为 11 个单元。隐含层单元的确定参考经验公式 $N=(\log_2{}^{N_1}+\log_2{}^{N_2})/2$，来确定初始网络结构中隐层神经元的个数，然后在仿真试验中逐步调整，从而确定最优隐层神经元个数。其中：N_1为输入神经元个数，N_2输出神经元个数，N 为隐层神经元个数。因此，隐含层单元数初步定为 4。

根据以上分析，BP 神经网络人群食物摄入量模型如下所示：



5.2.2 数据处理

（1）我们知道，用建立的人群食物量摄入模型去估计不同情况下人群的食物摄入量，其结果应该是一个范围区间，而不应该是一个精确的确定值，也就是说，当确定了地区、性别、年龄、季节、劳动强度和经济收入的情况下，用估计模型估计人的各类食物摄入量应该是一个合适的范围，而不能精确的说就是多少，这样才能增加估计模型的准确性和通用型。因此，我们必须对神经网络的输出进行适当的改进。通过对统计数据的分析我们得知，当确定了地区、性别、年龄、季节、劳动强度和经济收入的情况下，人对各类食物的摄入量基本服从正态分布，给定上述条件已蛋类为例如下图所示：

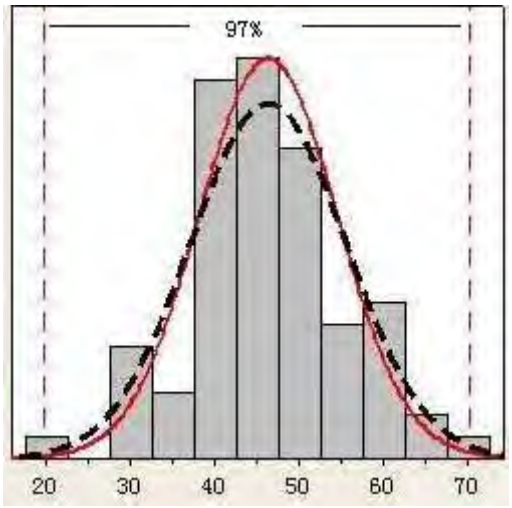


图 5-2 某种确定情况下人群蛋类摄入量分布图

上述选取了 97%的人群，在具体计算过程中，可以根据需要确定阀值（小于 100%），这样，只要神经网络的输出符合阀值确定的范围，即可认为是合理的食物摄入量，如果偏离，则进行误差累加，如果总误差平方和大于确定的标准则进行下一步网络训练，直到网络训练的结果小于确定的误差为止。

（2）为了让计算机区分不同的情况，使用分层确立数据向量的方式，以地区、年龄、季节、劳动强度、经济收入 6 个向量为第一层，由于一层变量自身属性不同，对其分类也不同，所以二层向量的个数不尽相同，但最多四个具体分类见上文。为达到计算机实现目的，需要将各向量用二进制数表示。二级向量共 20 个，用 5 位二进制数即可表示，为了便于理解，我们使用 8 二进制数表示，前 4 位表示第一层向量，后 4 位表示第二层向量。即将六个变量：地区、性别、年龄、季节、劳动强度和经济收入视为一向量：（地区，性别，年龄，季节，劳动强度，经济收入），用符号表示为：$(X_1,X_2,X_3,X_4,X_5,X_6)$。其中，$X_1$的取值为：$(00010001)$，$(00010010)$，$(00010100)$，$(00011000)$依次代表北一地区、北二地区、南一地区和南二地区。其余取值原理相同，依次用向量加以区分，这里不再赘述。向量确定如下表：

表 5-1 标量向量转化表

向　　量　　列　　表					
地区	性别	年龄	季节	收入	劳动强度
北一 00010001	男 00100001	少年 00110001	春 01000001	低 01010001	低 01100001
北二 00010010		青年 00110010	夏 01000010	中 01010010	中 01100010
南一 00010100	女 00100010	中年 00110100	秋 01000100	高 01010100	高 01100100
南二 00011000		老年 00111000	冬 01001000		

5.2.3 基于粗糙集的初始权值确定方法

粗糙集作为一种新兴的处理不精确、不确定与不完全数据的新数学理论，属性约简是其进行数据分析的主要工具，它能在保持知识库分类能力不变的条件下，通过属性约简，剔除冗余信息，导出问题分类和决策规则。此外，基于粗糙集的推理过程是完全由

数据本身决定的，无需要提供问题所需处理的数据集合之外的任何主观先验信息，更具客观性。这是粗糙集与其它不确定推理理论的最主要区别，也是最大的优点。

（1）基于粗糙集理论的实例知识表达

根据粗糙集知识表达系统，一个实例库（这里，实例也是知识的一种表达方式)可通过决策表的形式进行描述。该知识表达系统S可表示为

$$S=(U,R,V,f) \qquad \text{（公式 5-4）}$$

其中，U为非空的有限论域，表示实例库中所有实例的集合；$R=C\cup D$是实例特征属性的集合，子集C和D分别称为条件属性和结果属性，$C\cap D=\Phi$；$V=\sum_{a\in R}V_A$表示属性值的集合，V_a表示属性$a\in R$的取值范围。$f:U\times R\to V$是一个信息函数，它指U中每一对象（实例）的属性值，对$\forall x\in U,a\in R$，存在$f(x,a)\in V_a$。这样定义的知识表达系统可以方便地用关系数据表来实现，其中列表示属性，一个属性对应一个等价关系，行表示对象（实例），一个表可以看作是被定义的一族等价关系。

（2）粗糙集理论的相关定义

定义1 对于属性子集$B\subseteq R$，B的不可分辨关系定义为：

$$IND(B)=\{(x,y)\in U^2:\forall a\in B,f(x,a)=f(y,a)\} \qquad \text{（公式 5-5）}$$

其描述的是表格中两行数据对应属性集B的值域具有不可分辨的特性。易知，$IND(B)$满足自反性、对称性和传递性，一个等价关系。关系$IND(B)$构成了U的一个划分，用$U/IND(B)$表示，可简记为U/B。

定义2 对于每个对象子集$X\subseteq U$和不可分辨关系B，X的下近似集定义为：

$$B_{-}(X)=\cup\{E_i\mid E_i\in U/IND(B)\wedge E_i\in X\} \qquad \text{（公式 5-6）}$$

下近似集$B_{-}(X)$是由那些根据知识R判定肯定属于X的U中元素组成的集合。

定义3 设$P,Q\subseteq R$则属性P相对于属性Q的正域定义为：

$$POS_p(Q)=\sum_{x\in U/Q}P_{-}(X) \qquad \text{（公式 5-7）}$$

Q的P正域是U中所有根据分类U/P的信息可以准确地划分到关系Q的等价类中去的对象集合.

定义4 属性的约简，对于给定的知识表达系统$S=(U,R,V,F)$，设R是一个等价关系族，$a\in R$，如果$IND(R)=IND(R-\{a\})$，则称a在R中是可以被约去的属性。

（3）基于粗糙集的实例特征属性权重确定方法

一般来说，粗糙集理论为处理离散属性提供了一个很好的工具，但它不能直接处理连续属性，而现实中大多为连续数据，必须对这些数据进行离散化处理。目前，连续属性离散化的方法主要有：领域知识法、动态层次聚类法、布尔推理法以及统计检验法等。具体采用哪种方法及多大的离散粒度，应视隋况而定。

属性离散化后，可以得到离散数据决策表$S=(U,C\cup D,V,f)$，对于

$\forall a_k \in C$（$k\in[1,m]$，m为条件属性的个性），属性a_k的重要度$W_D(a_k)$可由下式进行计算：

$$W_D(a_k)=\left(card(POS_C(D))-card(POS_{C-\{a_k\}}(D))\right)/card(U) \qquad \text{（公式 5-8）}$$

上式中$card(X)$表示集合X的基。显然，$W_D(a_k)\in[0,1]$，$W_D(a_k)$越大，表明属性a_k对决策越重要。

当且仅当$W_D(a_k)>0$时，$a\in C$是条件属性子集C的一个必要属性。若某个

$W_D(a_k)=0$，说明a_k为冗余特征属性，应将其约去，然后再对属性约简后的决策表重新按(6)式进行特征属性重要度计算，直到所有 $W_D(a_k)>0$ 为止。最后按下式进行规范化处理，得到各属性的权重。

$$W_{(a_k)}=W_D(a_k)/\sum_{a_k\in C}W_D(a_k) \qquad \text{（公式 5-9）}$$

5.2.4 基于遗传算法的模型优化算法设计

传统的 BP 神经网络算法有不变学习速度算法和可变学习速度算法，众所周知，这两种算法易于操作，但同时也存在例如多次函数盲区、多个局部极小点等不可避免的缺点。为此，我们设计一种基于遗传算法的 BP 神经网络最优权值计算方法。

遗传算法 GA (Genetic Algorithm)，本质上是一种不依赖具体问题的直接搜索方法。其基本思想是：每一物种个体的基本特征会被后代继承，但后代又会产生一些异于父代的新变化，即在环境变化时，只有那些适应环境的个体特征才能保留下来。这些个体特征保存在细胞中，并以基因形式包含在染色体内，每个基因有特殊的位置并控制某种特殊性质。由于每个基因产生的个体对环境具有某种适应性，通过基因突变和基因杂交个体就可以产生更适应于环境的后代，这样经过存优去劣的自然淘汰使得适应性高的基因结构得以保存下来。

遗传算法 GA 把问题的解表示成“染色体”，在执行遗传算法之前，给出一群“染色体”，也就是假设解。然后，把这些假设解置于问题的“环境” 中，并按适者生存的原则，从中选择出较适应环境的“染色体”进行复制，再通过交叉，变异过程产生更适应环境的新一代“染色体”群。这样一代一代地进化，最后就会收敛到最适应环境的一个“染色体”上，它就是问题的最优解。

遗传算法的特点：

(1)遗传算法从问题解的集开始搜索，而不是从单个解开始，覆盖面大，利于全局择优；

(2)遗传算法求解时并不需要问题导数等其它直接与问题相关的信息，易形成通用算法程序；

(3)遗传算法的初始串集本身就带有大量与最优解甚远的信息，通过选择、交叉、变异操作这一强烈的并行滤波机制使得遗传算法有极强的容错能力；

(4)遗传算法中的选择、交叉和变异都是随机操作，而不是确定的精确规则；

(5)遗传算法具有隐含的并行性。

首先对符号规定如下：BP网络的初始权值：$W(W_1,W_2,\cdots,W_t)$，其中 $i=1,2,3,\cdots,m$；BP网络的输入值$X(X_1,X_2,\cdots,X_i)$，其中$i=1,2,3,\cdots,n$；BP网络的期望输出：$Y_i(Y_1,Y_2,\cdots,Y_i)$，其中$i=1,2,3,\cdots,k$；适应度函数： FitFunction；交差率：P_c；变异率：P_m。

该算法的基本思想是对随机产生的初始权值构造初始种群，按照适应度函数值对种群进行优良种的选择，淘汰适应度小的种群，形成新一代种群；对新一代种群根据P_c 进行个体的交叉，以产生带有优良信息的新个体，根据P_m进行个体的变异，以便在种群老化的时候产生新的个体，避免个体的早熟。然后对新一代种群进行适应度计算，直至最优适应度，这时也就找到了最优的初始权值，再利用得到的最优权值进行BP算法。

具体算法为：

(1)构造初始种群

设定种群规模，令其为N。利用随机方法产生初始权值种群$W^1\left(W^{11},W^{21},\cdots,W^{i1}\right)$，其中$1=1,2,3,\cdots,N$； $i=1,2,3,\cdots,m$。即构成了$m*N$的初始权值矩阵，问题的最优解将通过这些初始假设解进化而求出。

(2)适应度函数的确定（FitFunction）

适应度函数取BP算法的误差代价函数：$f\left(W^1\right)=\frac{1}{2}\left(X_i^m-Y_i\right)^2$。$f\left(W^1\right)$给出适应度的定义越小，称为适应度越大。那么最大适应度的求解过程即为对求最小值的过程。

(3)对种群进行选择（Selection）

对种群中具有良好遗传信息的个体（具有较好适应度的个体），在下一代进化种群中所占有数可通过下式计算出：

$$P_s=\left(1-\frac{f\left(W_i^1\right)}{\sum_{j=1}^{N}f\left(W^1\right)}\right)\bullet N$$ （其中P_s为选中W^1为下一代个体的次数）

上式反映出种群中适应度较高的个体，繁殖下一代的数目较多；适应度较小的个体，繁殖下一代的数目较少，甚至被淘汰。通过这种进化策略，就产生了对环境适应能力较强的后代，即选择出与最优解较接近的中间解。

(4)对种群进行交叉（Crossover）

在选出的下一代种群中，随机地选择两个个体W^i与W^j，按交叉概率P_c在选中的位置进行相同位置的交换。这个过程反映了随机信息交换，目的在于产生新的基因组合，即产生新的个体。

(5)对种群进行变异（Mutation）

在交叉后的种群中，以变异概率 P_m 对随机选取的一些个体按照生物遗传中基因变异的原理，对选中的个体的某些位执行变异（所谓变异，就是对欲执行变异的个体的某个串的对应位求反）。P_m 的取值应符合生物变异极小的规律。

(6)对新一代种群进行适应度计算，如找到最优权值 W 组结束；否则继续进行步骤(3)～(6)。

(7)将 W 组代入BP网络进行学习。

流程图如下所示：

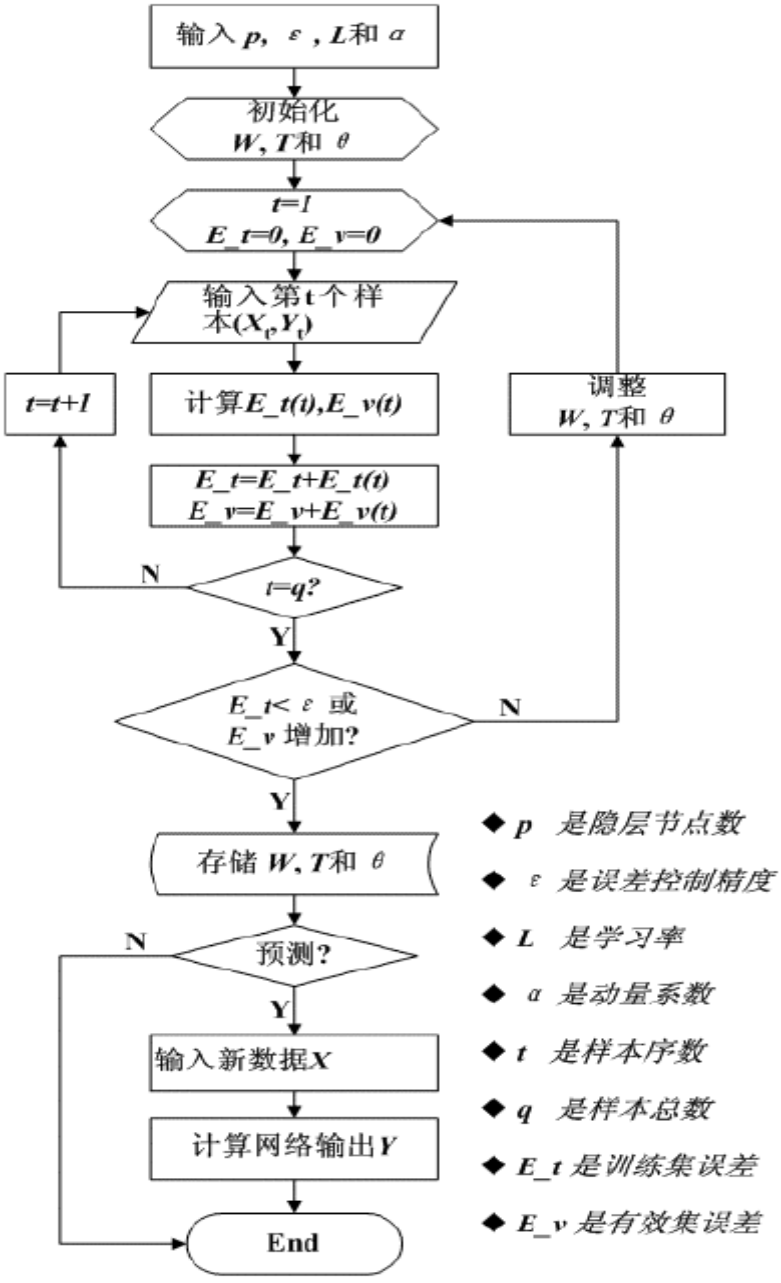


图 5-3 算法流程图

5.2.5 基于 Matlab 的人群食物摄入量神经网络模型仿真

采用 Matlab 神经网络工具箱，用 2000 个样本对网络进行训练，具体实现如下：

(1)根据基于粗糙集确定的神经网络的权值和阀值，定义模型的相关参数。该步主要用到的函数为 newff 函数，所定义的模型参数为网络训练过程中每隔多少步显示一次(net.trainParam.show)、训练的步数(net.trainParam.epochs)、误差指标(net.trainParam.goal)。

(2)定义输入向量 P 和目标向量 T，向量 P 由(2)中数据转换方法确定，向量 T 中国总膳食标准和统计数据确定。

(3)对神经网络进行循环训练，该步主要用到的函数为 net=train(net,p,t)。训练次数的确定以及停止训练的时间取决于误差 $\sum e^2$ （输出向量与目标向量误差平方和）是否达到满意的值。

(4)仿真输出，该步主要用到的函数为 a=sim(net,p)，其中 a 为输出向量。

应用传统算法和我们基于遗传算法设计的网络训练输出误差曲线如图 5-4 和图 5-4 所示。

根据 2000 个样本进行网络训练，下表给出了部分样本数据，全部训练样本见附录 3。

表 5-2 部分样本数据

地区	性别	年龄	季节	收入	劳动强度	谷类	蔬菜	水果	肉类	蛋类	水产	乳类	薯类	豆类	饮料类	酒饮类
黑龙江	男	29	秋	高	中	343	415	279	136	13	88	67	2	44	256	124
黑龙江	女	29	秋	低	中	330	342	135	15	69	21	115	46	50	369	0
辽宁	男	48	冬	低	高	593	560	245	58	109	67	55	36	30	236	43
辽宁	女	14	春	中	高	729	741	248	60	72	16	62	3	83	339	164
河北	女	32	秋	中	高	439	468	38	29	71	47	104	45	99	96	384
河北	男	83	春	中	低	346	404	73	102	73	14	117	51	39	122	150
河南	女	21	秋	低	高	637	565	297	22	29	45	76	47	100	384	327
河南	男	22	夏	高	高	514	691	227	58	6	143	121	126	58	347	0
宁夏	男	72	冬	中	低	493	383	91	32	44	26	54	9	35	121	233
宁夏	男	2	春	中	低	485	249	132	41	2	41	146	0	8	395	26
陕	女	53	夏	中	低	80	298	128	22	68	52	43	9	74	311	0

西																
陕西	女	19	秋	低	中	430	331	61	34	38	47	110	44	28	212	229
江西	男	40	春	低	高	725	606	145	28	22	0	129	26	71	163	280
江西	女	38	冬	中	中	650	374	281	90	4	35	113	60	38	326	0
上海	男	35	春	高	高	678	711	263	125	80	56	49	8	79	498	96
上海	女	32	秋	高	中	577	491	191	105	4	72	84	28	97	289	208
福建	女	70	夏	高	低	352	0	166	72	80	66	145	52	0	485	259
福建	男	50	冬	高	中	437	563	234	78	67	41	101	38	36	111	198
湖北	男	23	秋	低	中	359	100	121	2	46	7	77	34	27	172	275
湖北	男	18	春	高	高	503	549	294	112	29	54	118	45	29	344	184
四川	男	16	夏	低	中	313	314	93	57	56	67	47	55	79	351	370
四川	女	22	夏	高	高	704	491	152	75	62	41	45	2	93	133	27
广西	男	33	春	中	中	503	314	275	82	12	67	44	30	78	159	174
广西	女	6	秋	中	低	330	140	140	103	6	15	63	40	36	393	41

其中地区、性别、年龄、季节、经济收入和劳动强度根据 5.2.2 数据处理中阐述的方法进行转换，限于页面限制，上表中未给出转换后的向量。应用传统训练算法和基于遗传算法的神经网络训练算法对 2000 样本训练误差变化曲线如下所示：

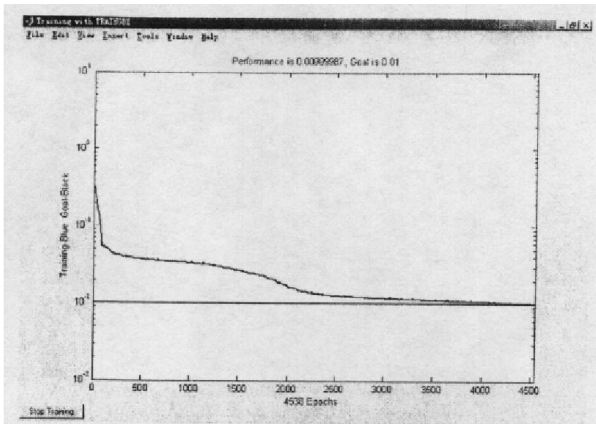


图 5-4 传统 BP 算法网络训练误差结果图

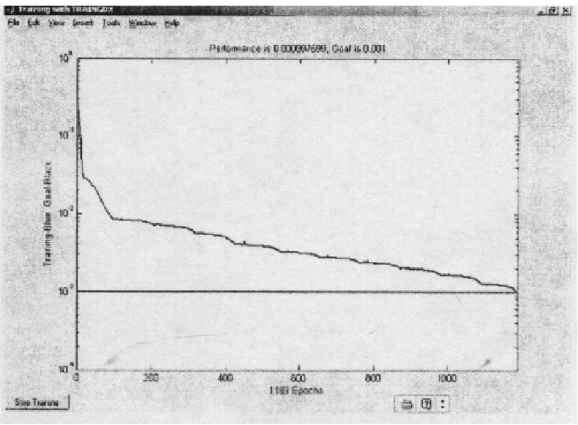


图 5-5 基于遗传算法网络训练误差结果图

传统 BP 算法训练到四千五百多次，误差才逐步接近要求，而基于遗传算法的网络

训练在一千三百次时误差迅速减小并满足要求，显示了基于遗传算法网络训练的优越性。

6 食物污染物分布模型的分析与建立

6．1 难点解决及建模基本思路

建立食物污染物分布模型是一项系统工程，涉及因素众多，加之我国地域辽阔，食品种类众多和许多不确定因素的影响，使得建立集准确性和通用型与一身的模型相当困难。正如文中所言，建立食物污染物模型面临三个主要难，解决思路为：

第一是如何解决如何在抽样率很低的情况下建立模型。我们认为这一问题主要是如何解决抽样率低的情况下使得样本数据具有典型代表性，从而使得即使在抽样样本相对总量很低的情况下能够代表总量的特征，这是问题的关键。解决这一问题我们认为应该创造性地设计抽样调查方案，即设计抽样调查方法，使用这种方法可以使得在抽样率低的情况下调查的数据可以代表总量的特征，从而利用抽样数据建立的污染物分布模型可以反映食物污染物的分布特征。因此，抽样调查方案的设计是这一问题的关键，在人群食物摄入量模型中，我们已经成功地建立了通用性很强的抽样调查方案，只要对部分细节进行稍加改动即可用于食物污染物的含量的抽样调查，鉴于数学建模是一项注重创造性地工作，此处我们不再花篇幅重复机械性的工作，请体谅。

第二是在无法获得详细数据的情况下如何提高模型的精度。这一问题的解决，除了尽量多的考虑季节、地区等不同情况对食物污染物分布的影响外，我们认为关键是不能局限于海量的数据和只将思路局限在数据的补偿上，我们认为在数据不完整、不详细的情况下提高模型建立的精度，应该**分两步走：**①首先对繁杂的相关数据进行分类，一类是对食物污染物分布有较大影响的数据；另一类是对食物污染物分布不具有影响或者影响很小的数据。对于前者数据重点考虑，而对于后者，即使数据再不完整再不详细也不影响模型的精度；②当然，对于第一类数据来说肯定也存在不完整不详细的数据，对于这一问题我们可以这样处理：根据已经获得数据分析食物污染物的分布随这部分数据的变化趋势。为了说明问题的方便，给出示意图如下所示：

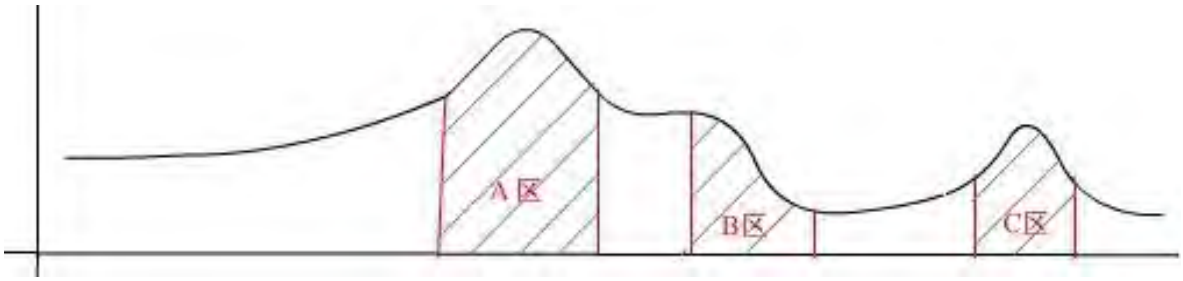


图 6-1 示意图

对变化趋势曲线进行分析，分为趋势变化急速区间和趋势变化平缓区间，对于趋势变化急速的区间例如图 6-1 中的 A、B、C 区，说明这些区间范围内的的数据对食物污染物分布模型影响较大，是重点考虑的数据；对于趋势变化平缓的区间，说明这一区间的数据对食物污染物分布影响较小，即使不完整也不会对模型精度造成较大的影响。经过这两步处理，可以使得在食物污染物分布模型建立的过程中大大降低对数据完整性和详细性的要求。

第三是在偶然性抽查本身就很低的情况下如何将未检出这部分信息加以利用，改变以往的未检出全部当成零来计算的现状，估计随机变量的整体分布。这一难点的解决首先要**认清两点**：①例行检查的次数和抽调样本数目远远大于偶然性检查，相对于偶然性

检测来说，例行检查抽调的样本是大样本，偶然性检测抽调的样本是小样本；②在偶然性检测中，抽调的样本中大部分食物是安全的，即相当于例行检测中的“未检出”，少部分食物是不安全的，既相当于例行检测中的统计各污染物含量的样本，为了下文说明问题的方便我们暂定为“”检出”。接下来主要进行**两方面的工作**：①是拟合或者估计例行检测数据和偶然性检测抽调样本数据中“检出”部分样本的污染物的分布，并总结其特征。②是利用例行检测的大样本验证偶然性检测抽调小样本的普遍代表性。基本原理是如果小样本的分布特征符合大样本的分布特征，则认为小样本具有普遍代表性。具体说就是将第一项工作得出的例行检测数据分布和偶然性检测中的“检出”样本数据分布进行相似性验证或者关联分析，如果两个分布具有明显的相似性则可以认为偶然性检测抽调的样本具有普遍代表性。同理，既然偶然性检测的样本具有普遍代表性，则其如果在例行检测中本属于“未检出”的样本数据亦具有普遍代表性，这部分在例行检测中本属于“未检出”的样本数据在偶然性检测中进行了详细的记录。因此，我们同样可以利用这部分分数据估计整个例行检测中的“未检出”，从而获得整个例行检测食物污染物分布，这样就获得了例行检测的完整数据。进一步可以进行更大样本的估计，并依次进行下去，可以获得任意大样本的食物污染物分布，从而可以比较准确的估计出食物污染物含量这个随机变量的整体分布。

6．2 食物污染物分布模型建立

此模型的建立主要包括三步：第一步是建立一个数据完整情况下的分布拟合模型；第二步主要工作是检验两个分布相似程度；第三步是基于上两步分析的结果估计食物污染物含量这个随机变量的整体分布：①用抽样调查方案进行例行检测和偶然性监测获得样本数据；②按问题分析中对不完整、不相信数据的处理方法对样本数据进行处理，以提高后文模型建立的精度；③估计食物污染物含量这个随机变量的整体分布。第一步的目的是得到例行检测中各类食物污染物的分布特征及曲线和偶然性检测中“检出”（上文已对”检出”定义）食物污染物分布特征及曲线。第二步的目的是检验这两个分布的相似度，如果相似度达到某一要求，则说明偶然性监测具有普遍性。可根据本在例行检测中属于“未检出”的这部分偶然性检测中的样本数据去估计例行检测中的“未检出”。第三步的目的是估计整个大样本的食物污染物含量这个随机变量的整体分布。

6.2.1 食物污染物分布的拟合预测模型

污染物分布模型要解决的问题是，根据农药、化工等污染行业的污染物排放数据和食品卫生安全监测部门日常对水、农贸市场和大宗食品中污染物的抽查数据以及进出口口岸的检测数据，来估计各类食物中各种污染物的含量。用数学语言严格地描述就是要设法根据随机变量取值大于某一数值的部分样本数据再加上其他可以利用的信息估计出这个随机变量的整体分布。我们首先分析处理数据，可以根据问题一中多变量多层混合抽样调查方案，抽取带有典型特征的抽样样本，并在此基础上针对几种危害面广、后果严重的污染物分别进行数据统计分析处理，得到几组关于各污染物和样本统计量的相关数据，结合这些数据运用回归拟合分析理论建立数学模型，然后通过 Mathmatica 数学软件编程，拟合得到关于各自污染物对不同食品的函数分布曲线。

回归分析是一种处理变量与变量之间关系的数学方法，如在各种实验过程中获取的观测数据以及对数据的分析整理、统计预报中的预测、经验公式中的参数确定等等，常常用到各种统计方法。应该指出的是，函数与相关虽然是两种不同类型的变量关系，但是它们之间并无严格的界限。一方面，如上所述，相关的变量之间尽管没有特定的关系，但从特定的统计意义上来看，它们之间又可能存在着某种确定的函数关系。本模型中我们通过回归拟合的方法，利用六种不同污染物类型在不同食品样品量分布的统计样本数据，拟合出反映各种不同污染物在不同食品中的分布函数表达式，从而顺延推理得到反

映各类食物中各污染物含量。

利用回归分析方法拟合不同类食物中各种污染物的含量与样本数量之间的函数关系，第一步，是要确定相关数据的回归方程，一般性多项式回归方程为：

$$y = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n \qquad \text{（公式 6-1）}$$

其中，$a_0, a_1, a_2, ..., a_n$ 分别为自变量体现不同类食物中各种污染物的含量随因变量抽样样本拟合曲线波动的波动系数。

通常利用最大似然估计法求取未知参数：对于 n 次多项式方程，有似然函数：

$$L(a_0,\ a_1,\ a_2 \cdots\cdots a_n) = \prod_{i=1}^{n} f(y_i) = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y-u(x_i)]^2} \qquad \text{（公式 6-2）}$$

要使 L 值取最大，应使上式指数中的平方和 $s = \sum_{i=1}^{n}[y_i - u(x_i)]^2$ 最小。换句话说，就是要分别求 s 对 $a_0, a_1, a_2, ..., a_n$ 的偏导数，令它们等于零，于是得方程组：

$$\begin{cases} \sum_{i=1}^{n}[y_i - u(x_i;\ a_0, a_1, ..., a_n)]\frac{\partial}{\partial a_0}u(x_i;\ a_0, a_1, ..., a_n) = 0 \\ \sum_{i=1}^{n}[y_i - u(x_i;\ a_0, a_1, ..., a_n)]\frac{\partial}{\partial a_1}u(x_i;\ a_0, a_1, ..., a_n) = 0 \\ ..... \\ \sum_{i=1}^{n}[y_i - u(x_i;\ a_0, a_1, ..., a_n)]\frac{\partial}{\partial a_n}u(x_i;\ a_0, a_1, ..., a_n) = 0 \end{cases} \qquad \text{（公式 6-3）}$$

从而得到自变量体现不同类食物中各种污染物的含量随因变量抽样样本拟合曲线各项波动的波动系数 $a_0, a_1, a_2, ..., a_n$；具体求解过程我们利用数学软件 Mathamatic 求解求得各未知参数，则针对不同食品的各污染物分布数据与样本统计量之间关系的回归拟合预测模型完成。

本题中由于污染物分布模型涉及到的污染物种类较多（我们主要选择了六种危害面广、后果严重的污染物展开分析），且食品包含谷类等十一种食品类型，还要针对不同地区、不同季节展开数据采样调查，数据量的拟合预测工作将达到 1056 组【1056=11（食品种类）×6（污染物类型）×4（季节）×4（地区）】回归拟合曲线，计算量太大，所以这里我们随机抽出一组数据展开回归拟合预测分析，以北一区、春季、谷类含铅量的污染物为基础，建立适用于其它 1056 组数据回归拟合的示范回归拟合模型如下：

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 + a_6 x^6 \qquad \text{（公式 6-4）}$$

用示意图反映如下：

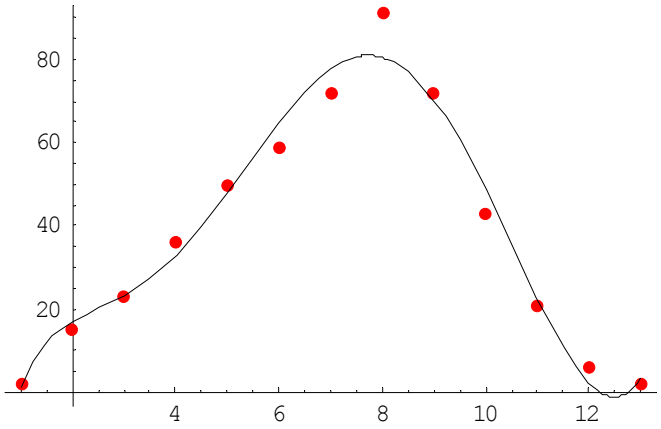


图 6-2 示意图

得到相应的函数拟合方程式为：

$y = 42.2867-63.9824x+27.5637x^2- 2.6041\ x^3-0.1529x^4+0.0311x^5-0.0011\ x^6$

具体拟合程序见附录 1。

6.2.2 对拟合数据的均方差检验分析

通过 Mathmatic 数学软件拟合，我们的拟合效果是否满意，还需要我们对拟合出来的数据进行均方差值的分析检验，观察拟合效果，对拟合出来的结果进行均方差分析如下：

用最小二乘法求拟合曲线时，首先要确定 P(x)的形式，这不单是数学问题，还与所研究问题的运动规律及所得观测数据(x_i, y_i)有关，通常是先用给定数据描图，确定 P(x)的类型，这种确定并不是一次就能完成，往往要经过确定几个类型后，通过实际计算比较均方误差，并选择均方误差较小的拟合函数，一般说来拟合函数中待定系数越少越好。

均方误差的计算公式：

$$S = \sqrt{\sum_{i=0}^{m}[y_i - p(x_i)]^2} \qquad \text{（公式 6-5）}$$

表 6-1 拟合出来的铅元素污染物统计分布数据均方差值检验

点 / 样本	1	2	3	4	5	6	7	8	9	10	11	12	13
y_i	2	15	23	36	50	59	72	91	72	43	21	6	2
$p(x_i)$	1.5	16.	23.1	34.6	48.8	62.	74.1	89.3	70.4	45.	22.1	4.2	3.1
$\lvert y_i - p(x_i)\rvert$	0.5	1	0.1	1.4	1.2	3	2.1	1.7	1.6	2.	0.9	1.8	1.1

计算得到，拟合出来铅元素污染物统计分布数据的均方差值 s=3.16035，满足均方差要求，认为拟合效果较好的反映了北一区、春季、谷类含铅量的污染物概率密度分布函数，可以推广。

6.2.3 两分布相似检验

经过食物污染物分布的拟合预测模型，我们会得到将要检测的两个分布，假设为 X 和 Y，密度函数为 $f1$ 和 $f2$，检测区间为 (a,b)，检测区间的取值要根据食物中某种污染物含量的大致变化范围确定。对两个分布密度函数在区间 (a,b) 上取点求值，对同一个点两个密度函数的对应值进行取差求平方，定为在这一点上两个分布密度函数的误差，然后求其所有误差平方和 $\sum\theta$，这个和小于某一要求的临界值 δ，则认为两分布相似。

具体操作如下：

（1）取 $(x1,x2\cdots,xn)\in(a,b)$；

（2）根据实际需要确定临界值 δ，不同情况对 δ 的要求不同，根据具体情况确定；

（3）计算误差平方和 $\sum\theta=\sum_{i=1}^{n}\left(f1(xi)-f2(xi)\right)^2$

（4）将误差平方和 $\sum\theta$ 与临界值 δ 进行比较，验证两分布的相似性。

6.2.4 食物污染物含量分布

这一部分中的阐述以谷类为例说明食物污染物含量分布的确定方法。根据抽样调查方案调查样本，分为例行检测和偶然性监测抽调两种，部分样本数据如表 6-2 和表 6-3 所示，限于篇幅限制，这里给出了部分数据样本，详细数据样本见附录 4。回归预测例行检测和偶然性监测中的“检出”样本污染物含量分布密度函数，根据 6.2.3 中相似检验的方法进行分布密度函数的相似性验证，如果相似则相信偶然性检测数据的普遍性，如果不相似则重新进行偶然性检测。

表 6-2 例行检测数据样本

北一区春天谷类例行检测数据						
	铅	镉	砷	汞	有机氯	有机磷
1	272	160	238	445	616	1358
2	253	197	296	352	601	1666
3	208	170	251	411	700	1026
4	216	184	258	439	536	1454
5	503	100	243	114	557	1004
6	283	248	242	312	624	1794
7	290	158	280	262	698	1639
8	385	126	216	275	590	2423
9	351	157	260	518	851	1444
10	450	188	241	267	544	1183

11	393	157	253	447	554	1674
12	347	174	401	469	503	1281
13	398	121	264	630	512	1359
14	256	231	257	350	500	1358
15	329	175	288	551	577	1007
16	298	110	290	562	989	1356
17	332	152	280	293	610	2094
18	277	128	301	518	665	1648
19	249	115	224	447	652	1122
20	355	115	252	592	637	1144
21	345	196	296	478	503	1087
22	270	133	294	305	672	1112
23	384	103	416	259	695	1123
24	226	200	230	179	587	1579

表 6-3 偶然性检测数据样本

北一区春天谷类偶然性检测数据						
	铅	镉	砷	汞	有机氯	有机磷
1	119	6	72	53	171	948
2	101	3	303	124	49	733
3	63	20	29	35	381	335
4	56	14	44	69	160	612
5	137	19	158	118	563	282
6	89	20	100	157	492	72
7	199	44	129	63	215	1725
8	62	8	199	95	90	959
9	10	17	95	44	335	704
10	139	8	55	94	241	697
11	3	16	155	100	406	855
12	365	19	21	219	260	97
13	91	116	561	187	419	660
14	288	15	51	257	29	1648
15	9	5	61	57	125	539
16	258	11	34	88	211	229
17	25	12	80	91	242	408
18	71	17	144	40	238	810
19	335	8	83	89	342	92
20	87	220	385	173	328	803
21	56	16	57	208	311	436
22	204	5	195	28	79	765
23	52	4	51	24	47	183
24	89	14	208	91	178	263

有底纹的数据表格表示偶然性监测中某种污染物含量超标。

食物中各污染物含量限量指标如下表所示：

表 6-4 食品中污染物限量指标

食品	铅	镉	汞	砷	有机氯	有机磷
谷类	200	200	200	150	500	1000
蔬菜	100	100	100	50	500	500
水果	100	50	100	50	500	500
肉类	200	100	500	50	5000	100
蛋类	200	50	500	50	1000	100
水产	500	100	5000	100	1000	100
乳类	50	无标准	100	50	1000	100
薯类	200	100	100	200	500	100
豆类	200	200	无标准	100	500	100
饮料类	50	无标准	无标准	200	500	100
酒类	200	无标准	无标准	50	200	100

以上标准是根据《中华人民共和国国家标准》GB 2762－－2005 食品中污染物限量中分别使用条款 GB/T 5009.12、GB/T 5009.15、GB/T 5009.11、GB/T 5009.17 规定的铅、镉、汞、砷的污染物限量标准。有机氯和有机磷的限量标准根据相关研究人员根据对人体的危害程度等判断，其准确性有必要进一步验证。为便于处理数据，将以上标准单位作如下定义：

表 6-5 每千克食品中各污染物含量单位

各污染物含量单位确定					
铅	镉	砷	汞	有机氯	有机磷
ug/kg	ug/kg	ug/kg	0.1ug/kg	0.01ug/kg	0.01ug/kg

经过多次抽测，得到符合要求的偶然性监测样本数据，求得例行检测和偶然性监测中的“检出”样本污染物含量分布密度函数如下表所示：

表 6-6 污染物含量分布密度函数

污染物	例行检测食物污染物含量分布密度函数	偶然性检测中“检出”食物污染物含量分布密度函数
铅	$f_1(x)$=-248.840+ 310.976 x-138.549 x^2 +30.416x^3-3.360x^4+0.179x^5 -0.004x^6	$f_1'(x)$=-236.398+295.427x- 131.622x^2+28.896x^3-3.192x^4+ 0.170x^5-0.004x^6
镉	$f_2(x)$=-52.986+ 92.876x - 49.547x^2 + 12.964x^3 - 1.591x^4 + 0.089x^5 - 0.002x^6	$f_2'(x)$=-45.568+79.873 x-42.610 x^2 +11.149x^3-1.368x^4 + 0.077x^5 -0.002x^6
有机磷	$f_3(x)$=-66.322+ 118.9842x -63.662x^2 + 16.416x^3 - 2.016x^4 + 0.115x^5 - 0.002x^6	$f_3'(x)$=-64.332+ 109.605 x-61.752 x^2 + 15.924x^3-1.956x^4 + 0.112x^5 -0.002x^6

有机氯	f_4（x）=-58.238+99.338x-51.327x^2 + 13.017x^3 - 1.554x^4 + 0.085x^5 - 0.002x^6	f_4'（x）=-54.747+ 93.378x-48.247 x^2 +12.236x^3-1.461x^4 + 0.080x^5 -0.002x^6
汞	f_5（x）=-59.993+ 104.6551x - 54.239x^2 + 13.799x^3 - 1.673x^4 + 0.094x^5 - 0.0012x^6	f_5'（x）=-54.594+ 95.236x-49.357 x^2 +12.557x^3-1.522x^4 + 0.086x^5 -0.001x^6
砷	f_6（x）= -44.007+ 78.292x-41.356x^2+0.833x^3-1.313x^4+0.072x^5 -0.001x^6	f_6'（x）=-38.726+ 68.017 x-36.393 x^2 + 0.733x^3-1.155x^4 + 0.063x^5 -0.001x^6

各种污染物含量两种检测密度函数相似度大小为：铅>有机氯>有机磷>砷>镉>汞。然后以如在例行检测中本属于“未检出”的那部分偶然性监测获得的样本数据估计例行检测中的“未检出”，估计食物整体污染物含量这个随机变量的整体分布曲线如下图组所示：

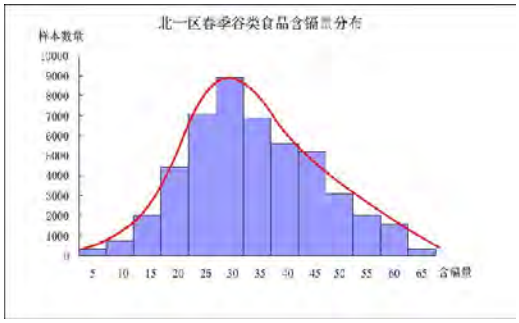


图 6-3 北一区春季谷类含镉量分布曲线

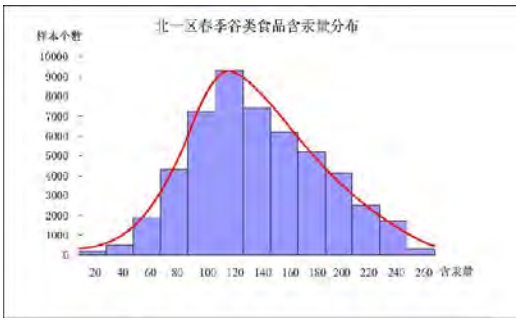


图 6-4 北一区春季谷类含贡量分布曲线

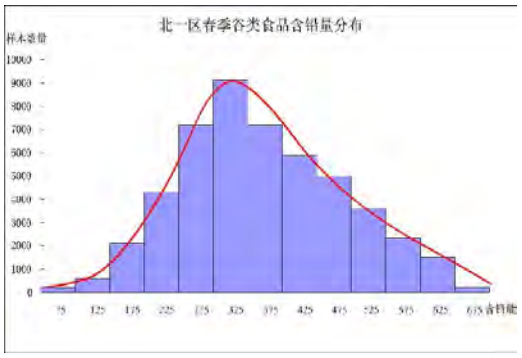


图 6-5 北一区春季谷类含铅量分布曲线

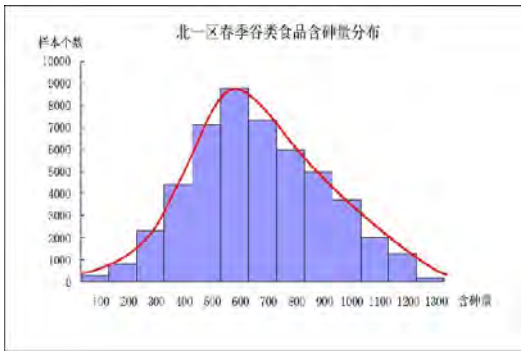


图 6-6 北一区春季谷类含砷量分布曲线

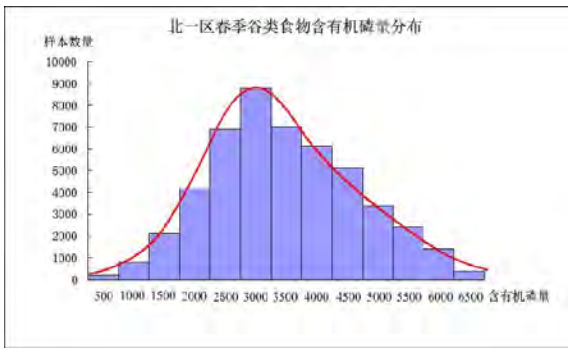


图 6-7 北一区春季谷类含有机磷量分布曲线

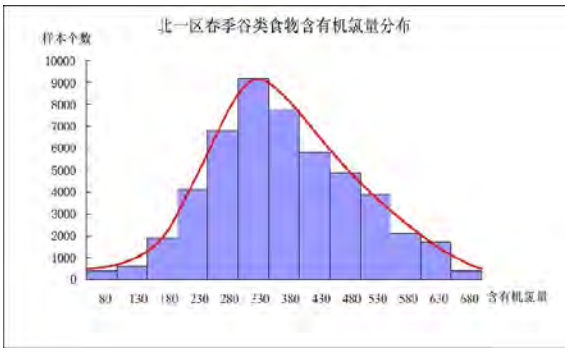


图 6-8 北一区春季谷类含有机氯量分布曲线

7 风险评估模型的分析与建立

建立风险评估模型，总的意图在于利用前两个模型的结果对全国、某个地区、某类食品的安全状况做出评价，对可能出现的食品安全事件给出预警。具体来说，是根据前两个模型所提供的数据，计算得出全国或某地区人群某些污染物每天摄入量的99.999%的右分位点，从而能够对某一时刻食品安全风险做出评估。如果这个右分位点的数值明显小于由食品卫生安全主管部门制定的、经过大量试验被证明是安全的标准，则我们就有比较充分的理由相信目前的食品卫生状况是安全的。用数学语言严格地表述，就是如果把每个人每天某种污染物摄入量看成是一个随机变量，则我们关心的不仅是它的均值，更关心的是它的99.999%的右分位点数值。

7．1 食物风险评估模型相关问题分析

关于风险评估模型，需要主要解决的问题有：

第一，题目指出该风险评估模型的输入都是抽样率很低的随机抽样的数据，而且这两批数据是不配套的，即人群食品摄入量模型中的调查对象极大可能不是污染物分布模型中被调查食品的消费者，如何根据上述两批结果建立风险评估模型是第一个难点。关于此问题，我们认为，虽然人群食品摄入量模型中的调查对象很可能不是污染物分布模型中被调查食品的消费者，但是作为一个地区、一个国家而言，人群食品摄入量模型中的概率分布总是可以反映一定的偏态分布规律，这就又返回到问题一中的抽样调查方案设计的问题，不再赘述。

题目给出的第二个难点是：两个模型的数据分类可能不配套，人群食品摄入量模型中的食品很可能远多于污染物分布模型中被调查食品或者两者的分类不完全一致，这个问题我们做规避性处理。因为总体是根据一定的目的和要求所确定的研究事物的全体，它是由客观存在的、具有某种共同性质的许多个别事物构成的整体，它代表考察对象的全体；个体是指某个单位或物体的其中一个，它是具有特殊性的、总体中的每一部分。许多的个体组成一个总体，总体中每个成员称为个体。我们的研究思路是考察总体属性中的一些共性问题，然后建立模型找出该共性问题之间的规律性东西，但是涉及个体属性中的某些特殊问题的研究，由于个体对象的非典型性和超独立性特点，决定了对该类问题的讨论毫无意义。所以风险评估模型的难点就转化到第三个难点的分析和建模上面，也即要求模型给出全体居民某项污染物摄入量的99.999%的右分位点，并且分析如何才能提高模型的精度。

在进行具体建模数据分析处理时，我们选取前两个模型所提供的典型数据，针对某一特定地区、特定季节、特定食品的特定污染物的数值，拟合分析得到该特定地区在特定季节对特定食品的特定污染物数据处理的密度分布函数，结合该密度分布函数，积分得到该特定地区在特定季节对特定食品的特定污染物摄入量风险检验的概率分布数据，按照99.999%的标准展开评估和分析工作。当然，这只是一组数据的概率分布模型，但是该模型可以推广到其它4×4×11×6=1156（四个大区，四个季节，十一项食品以及六项污染物）组类似摄入量风险检验评估模型中去，由于时间原因，该项工作我们选取北一区、春季、谷类食品关于铅含量摄入量的样品数据展开建模分析，以期起到抛砖引玉的作用。

7．2 食品中污染物摄入量的概率求取模型

问题一拟合得到北一区、春季、谷类食品关于铅元素的摄入量概率密度分布函数为：

f (x) = 42.2867-63.9824x+27.5637x^2- 2.6041 x^3-0.1529x^4+0.0311x^5-0.0011 x^6

在此基础上，建立积分模型如下：

$$P_q = \int_{-\infty}^{x} f(x)dx \qquad \text{（公式 7-1）}$$

其中，P_q 表示北一区、春季、谷类食品关于铅元素的摄入量的概率值。则：

$$P_q = \int_{-\infty}^{x} (42.2867-63.9824x+27.5637x^2-2.6041x^3-0.1529x^4+0.0311x^5-0.0011x^6)dx$$

（公式 7-2）

进而，可得到北一区、春季、谷类食品关于铅元素的摄入量概率，有

$$P_q=(42.2867x-31.9912x^2-9.1879x^3-0.651025x^4+0.03058x^5-0.0052x^6-0.0011x^7)\Big|_{-\infty}^{x}$$

（公式 7-3）

由于铅元素的摄入量在统计分布区间的取值自零开始，所以上式可转化为：

$$P_q=(42.2867x-31.9912x^2-9.1879x^3-0.651025x^4+0.03058x^5-0.0052x^6-0.0011x^7)\Big|_{0}^{x}$$

（公式 7-4）

这样一来，就有：

$$P_q=42.2867x-31.9912x^2-9.1879x^3-0.651025x^4+0.03058x^5-0.0052x^6-0.0011x^7$$

（公式 7-5）

依此类推，我们可以建立上述分析的 1156 组抽样样本数据的污染物拟合模型，也即建立起针对不同地区、不同季节的各种类型食品的污染物元素摄入量的概率密度分布函数，根据此概率密度函数，按照上述积分模型的思想即可得到保证绝大多数（99.99%以上）居民食品安全的风险评估数据库，不仅解决了题目关于风险评估模型中的居民某项污染物摄入量的 99.999%的右分位点问题，也即解决了该模型中涉及污染物分布呈现偏态分布的精度处理等相关理论问题。

7．3 模型三的算法

简单的说就是积分建模算法，具体步骤如下：

总体来说，就是求曲边梯形 $\{(x,y)\mid a\le x\le b,0\le y\le f(x)\}$ 的面积，可分为下列 4 步：

① 分割：把 $[a,b]$ 任意分成 n 个小区间 $[x_{i-1},x_i]$，$i=1、2、\cdots、n$, $\lambda=\max\{\Delta x_i\}$。

② 近似代替：$\Delta s_i\approx f(\xi_i)\cdot\Delta x_i$。

③ 求和：$s=\sum_{i=1}^{n}\Delta s_i\approx\sum_{i=1}^{n}f(\xi_i)\cdot\Delta x_i$。

④ 取极限：$s=\lim_{\lambda\to0}\sum_{i=1}^{n}f(\xi_i)\cdot\Delta x_i=\int_a^b f(x)dx$。

在这个过程中，关键的是以下两步：

无限分割：设想把 $[a,b]$ 无限细分，得到无数多个小区间。取其中一个典型区间 $[x,x+dx]$，对应地，得到 $[x,x+dx]$ 上的小曲边梯形。根据微分模型的几何意义，得面积

微元$d_s = f(s)dx$。

无限相加：把$[a,b]$上的无数多个小曲边梯形的面积微元无限相加，得$s = \int_a^b f(x)dx$。一般地，如果某一实际问题中所示量U符合下列条件：

1）U与一个变量x的变化区间$[a,b]$有关；

2）U对于区间具有可加性。即若把$[a,b]$分成许多部分区间，则U相应地分成许多部分量ΔU_i，且$U = \sum \Delta U_i$；

3）$\Delta U_i \approx f(\xi)\cdot \Delta x_i, (\xi_i \in [x_{i-1}, x_i], i = 1,2,\cdots,n)$，

那么，就可以用定积分求出

$$U = \int_a^b f(x)dx \qquad \text{（公式 7-6）}$$

这样，我们就得到了可用定积分解决实际问题的数学模型——微元法。不难看出，微元法的第一步就是把U无限细分成无数多个部分量ΔU，且$\Delta U \approx dU$，得到U的微元dU，这一步实际上就是把U微分；第二步是把这些ΔU无限相加，即在$[a,b]$上把dU累积起来，得$\int_a^b dU$，这一步实际上就是积分——累积微分。

7．4 模型的求解

该模型的求解，说到底就是一个简单的对概率密度分布函数积分求概率的问题。

由于P_q表示北一区、春季、谷类食品关于铅元素的摄入量的概率值。则有：

$$P_q = \int_{-\infty}^{x} (42.2867-63.9824x+27.5637x^2-2.6041x^3-0.1529x^4+0.0311x^5-0.0011x^6)dx$$

由于铅元素的摄入量在统计分布区间的取值自零开始，所以上式可转化为：

$$P_q = (42.2867x-31.9912x^2-9.1879x^3-0.651025x^4+0.03058x^5-0.0052x^6-0.0011x^7)\Big|_0^x$$

这样就有：

$$P_q = 42.2867x-31.9912x^2-9.1879x^3-0.651025x^4+0.03058x^5-0.0052x^6-0.0011x^7$$

令$P_q = 0.99999$，则计算得到七个数据，分别为：

$x1 = -8.83485$；　$x2 = -5.19733 - 3.80272i$；　$x3 = -5.19733 + 3.80272i$；
$x4 = 0.0240901$；　$x5 = 0.989689$；　$x6 = 6.74423 - 7.65376i$；
$x7 = 6.74423 + 7.65376i$

分析得到的结果，发现符合条件的取值有两个，分别为$x4 = 0.0240901$ ug；$x5 = 0.989689$ ug。这就是说，北一区、春季、谷类食品关于铅元素的摄入量在 99.999%的右分位点的数值是 0.989689ug，而 2005-01-25 发布的中华人民共和国国家标准《食品中污染物限量》第 58 页表 1 关于“食品中铅限量指标”指出，谷类食品的限量指标是 0.2mg，而 0.989689ug<<0.2mg，故而认为这个右分位点的数值明显地小于由食品卫

生安全主管部门制定的、经过大量试验被证明是安全的标准，所以我们充分相信目前北一区、春季、谷类食品的卫生状况是安全的，再加上我们选择抽样样本的典型性和独立性优势，所以我们也充分相信样本选择的谷类食品可以保证全国各地区、各年龄段更多居民的食品安全。

依此类推，我们可以建立上述分析的 1156 组抽样样本数据的污染物拟合模型，也即建立起针对不同地区、不同季节的各种类型食品的污染物元素摄入量的概率密度分布函数，根据此概率密度函数，按照上述积分模型的思想即可得到保证绝大多数（99.999%以上）居民食品安全的风险评估数据库，不仅解决了题目关于风险评估模型中的居民某项污染物摄入量的 99.999%的右分位点问题，也即解决了该模型中涉及污染物分布呈现偏态分布的精度处理等相关理论问题。

8 食品卫生安全保障系统设计及展望

8.1 食品卫生安全需求分析

人类的任何实践活动都是有目的的，这些目的是一种行为或作用的结果，统称为作用。任何一个作用包含两方面内容:主动实体、被动实体。被动实体接受作用并受影响而发生改变，当这一改变符合预期，认为是需求。被动实体用特性或状态来描述，通常被动实体受到作用后,部分特性发生改变或状态发生变化。据此定义需求为：被动实体的特性和状态的符合预期的改变或变化。

世界卫生组织警告，人畜共患疾病、食品恐怖主义、食品污染和食源性疾病等正在威胁全球的食品安全。特别是近几年，危及人类健康和生命安全的重大食品安全事件屡屡发生。如疯牛病、口蹄疫、禽流感、二噁英、苏丹红、瘦肉精、吊自块、农药残留、兽药残留等等，这些国内外重大的食品安全事件不仅影响到人民群众身体健康和生命安全，而且还严重影响经济的发展，有的甚至导致贸易纠纷，从而影响社会稳定。尽管我国建立了一系列法律法规，并由专业队伍进行食品安全监测和执法，使食品安全监管取得了一定的效果，但是食品在生产、加工、流通过程中仍存在诸多不良因素，危害着我国食品安全。这样看来，食品系统安全保障系统设计需求显然存在。

8.2 食品卫生安全保障系统功能需求分析

与被动实体相对应，主动实体发生作用于被动实体，使被动实体产生变化，这种机制称为功能。主动实体也可由特性来描述，通常一项功能取决于主动实体的部分特性。故功能定义为：主动实体部分特性的作用，这种作用能使被动实体发生符合预期的变化或状态的改变。功能因需求而存在，需求决定功能。

食品系统安全保障系统的功能，在于建立检测方法和开展监测研究，通过对各种消费者人群进行暴露评估，得出食品危险性评价的最终结果，发展危险性评估技术和建立食品污染监控计划，期望为国家掌握食源性疾病的变化趋势和制定食源性疾病控制对策提供重要依据。具体来说，就是：通过一个有别于一般的评价手段并在一定的原则下建立与健康声称相关的食品及食品成分的评判体系；评价现有的评估食物功能声称的体系；在现有的一般评判体系中筛选那些可被识别、合法化的数据，从而纳入到严谨的研究饮食与健康关系的评价体系中。

8.3 内部结构

食品系统安全保障系统的内部结构如图 8-1 所示：

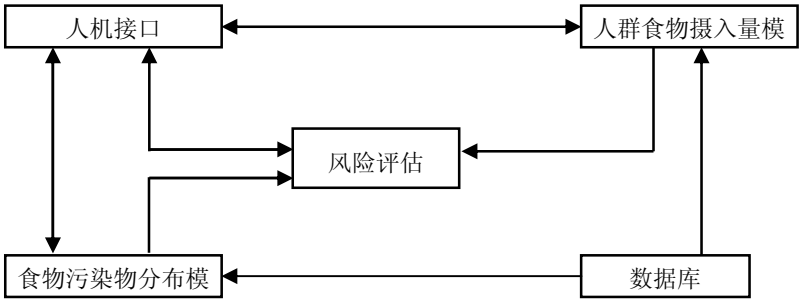


图 8-1　食物安全保障体系设计

8.3 系统开发及应用

我们初步对食品系统安全保障系统建立食物健康声称的系统开发临时标准如下：

(1)有声称的食品及食品成分必须遵循现行的法律体系。

(2)健康声称必须建立在完整的科学证据之上。经科学证实的评价体系是有用处的，但并非是必需的。

(3)一旦发布声称，必须注明此类产品的适用人群，比如：所有人群或某一特定人群。

(4)声称应主要建立在可观测的人类干扰试验的基础上。这应是一个与研究目的相符的合乎科学的实验设计，包括：

a．选定受试人群应在目标适用人群中具有代表性；

b．对实验组及对照组同样加以控制；

C．声称功能所表现出的效果需要有一定的持续性；

d．需要对受试人群的饮食习惯进行调查并加以控制；

e．对受试者的给药量应当与一般消费者的服用量相当；

f．需要建立相应的量效关系以决定消费者的最佳服用量；

g．其服用过程中相应的饮食要求同样应得到监控；

h．运用正确的统计方法。

(5)如经检测功能无法确认，应运用目前已有的科学评价方法进行研究。

(6)功能标记应科学、合理。方法学上的要求包括：①准确性和正确性；② 特异性及灵敏度；③可重复性。符合生物学规律，包括：①功能的产生可以用现有的生理学理论解释；②在实验中施以相关影响后机体能够做出及时的反应。

(7)在实验中通过正常的生物学途径，标记物发生反应并能产生统计学上的显著性的功能改善。

基于此标准的食品安全风险评估系统开发界面如图 8-2 所示：

图 8-2 食品安全风险评估系统开发界面

食品安全风险评估系统应用实例如图 8-3 所示：

图 8-3 食品安全风险评估系统应用实例

[参考文献]

[1] 彭祖赠，数学模型与建模方法，大连海事大学出版社，大连，1997
[2] 叶其孝，大学生数学建模竞赛辅导教材，湖南教育出版社，长沙，1993
[3] 张志涌，精通 MATLAB6.5 版，北京航空航天大学出版社，北京，2003
[4] 王小东等，一维下料优化的一种新算法，大连理工大学学报，第 44 卷，第 3 期，2004 年 5 月
[5] 张春玲、崔耀东，一维优化下料问题，桂林工学院学报，第 24 卷第 1 期，2004 年 1 月
[6] 黄崇斌，二维板材优化下料快速搜索法，计算机辅助工程，第 1 期，2000 年 3 月
[7] 林晓颖、王远，多目标优化下料问题的研究，哈尔滨师范大学自然科学学报，第 19 卷第 5 期，2003 年
[8] http://www.astrokettle.com
[9] G.Belov and G.Scheithauer, The Number of Setups(Different Patterns) in One-Dimensional Stock Cutting, www.math.tu-dresden.de/capad, September23,2003
[10] G.Belov and G.Scheithauer, Setups and Open Stacks Minimazation in One-Dimensional Stock Cutting, www.math.tu-dresden.de/capad, June24,2004
[11] 宋翔、聂义勇，无限制二维下料问题的改进动态规划算法，信息与控制，第 32 卷第 1 期，2003 年 2 月
[12]刘承平. 数学建模方法[M]. 北京：高等教育出版社，2002
[13]姜启源. 数学模型（第二版）[M]. 北京：高等教育出版社，1993
[14]张维迎. 博弈论与信息经济学[M]. 上海：上海人民出版社，1996
[15]苏金明，张莲花等. MATLAB 工具箱应用[M]. 北京：高等教育出版社，2004
[16] 李丽霞，王彤，范逢曦 BP 神经网络设计探讨[J]. 现代预防医学,2005,32(2);128-129
[17] 黄俊,张斌,鲁艺.基于改进 BP 网络的武器效能评估专家系统[J]. 电光与控制,2005,12(4) ;29-30
[18] Martin T.Hagan.神经网络设计[M].北京:机械工业出版社,2002,197-235
[19] 罗建军,杨琦.MATLAB 教程[M]. 北京:电子工业出版社,2005,85-104
[20] 中华人民共和国国家标准 GB 2762-2005 ,中华人民共和国和中国国家标准化管理委员会发布.
[21] 李金昌, 应用抽样技术[M]. 科学出版社 2006.
[22]（美）芬克（Fink A.）黄卫斌, 如何抽样[M]. 中国劳动保障出版社,2004.
[23] http://www.math.sjtu.edu.cn/Mathematica%BD%CC%B3%CC/ Mathematica 教程
[24] 刘宏志,陈惠京,王绪卿 1992年中国总膳食研究---农药残留[J]. 卫生研究1995 第 6 期
[25] 李筱薇,高俊全,陈君石 2000 年中国总膳食研究---膳食汞摄入量[J]. 卫生研究 2006 第 3 期.
[26]高俊全, 李筱薇,赵京玲 2000 年中国总膳食研究---膳食铅、镉摄入量[J]. 卫生研究 2006 第 6 期.
[27] 李筱薇,高俊全,王永芳，陈君石 2000 年中国总膳食研究---膳食砷摄入量[J]. 卫生研究 2006 第 1 期.

[28]赵云峰，李敬光，封锦芳等，2000 年中国总膳食中六六六和滴滴涕污染的溯源性分析[J]. 2004 第 4 期.
[29]张惠英，宁琦如等，宁夏居民膳食结构及营养状况调查分析[J].宁夏医学院学报，2005 第 2 期.
[30]周锟，浅谈抽样检验[J]. 维普资讯 http://www.cqvip.com.
[31] 赵云峰,吴永宁,王绪卿,高俊全,陈君石 中国居民膳食中农药残留的研究[J]. 中华流行病学杂志 2003 第 8 期.
[32]中华人民共和国总务部，出口商品技术指南－食品污染物、农残限量[M]. http://kjs.mofcom.gov.cn/table/cksp/3/wuranwu.pdf.
[33] 陈君石，高俊全,1992 年中国总膳食研究一化学污染物[J]. 卫生研究 1997 第 3 期
[34] 中华人民共和国外经贸行业标准 WM/T2-2004 药用植物及制剂外经贸绿色行业标准 2005.
[35] 唐晓纯 食品安全预警体系评价指标设计[J]. 食品安全维普资讯 http://www.cqvip.com
[36] 戴廷灿，李伟红，卢普滨，周海杰我国食品安全面临的问题及防范对策[J]. 江西农业学报 2003 第 2 期.
[37] http://hi.baidu.com/ahbbwf/blog/item/9c5ed38027f3e7d59123d98a.html.
[38] 管刚，食品安全综合评价研究 浙江大学 2007 年硕士研究生论文.
[39] 张馨，中国南北方居民一年食物摄人量的变化趋势研究[j]. 中国疾病预防控制中心营养与食品安全所，北京.
[41]何丽，赵文华，张馨，等. 中国南北方居民一年中营养素摄入量的变化趋势研究. 卫生研究. 2004，33(6)：694. 697.
[42] 中国营养学会编著. 中国居民膳食营养素参考摄入量(ChineseDRIs). 北京：中国轻工业出版社，2000. 96—99.
[43] 中国膳食指南专家委员会主编. 中国居民膳食指南文集. 北京：中国检察出版社，1999
[44] 赵文华，由悦，张馨，等. 中国不同“菜系”地区中老年人的膳食模式及食物摄入量研究. 卫生研究，2002，31(1)：34. 36.

附录

附录 1 模型 1 的 c 实现

```
float w[Ni+No+Nm],w[Ns][N];
int flag;
/*all functions*/
int myrand()
float Fitfunction(float array[],int n)
void Swap(w1[Ns][N],i,j)
void Mutation(w1[Ns][N],i,j)
float Sigmoid(float a)
float FrontCal(float a,float b,aray[],int n)
main()
{float x[2],s1[2],s2[4],f1[2],f2[4],d1[2],d2[4],t[2],s,f,d,test;
int i,j,N,Fw[Ns],AllFit,BpTime,y,Ttime,Rtime;
N=Ni+No+Nm;
float w1[Ns][N],w1[N];float w1[Ns][N]
int Ps[N],t,n,Gtime;
/*GA program*/
for(i=0;i<Ns;i++)
for(j=0;j<N;j++)
W[i][j]=pow(-1,myrand()%2)*myrand()%A;
AllFit=0;
while(flag!=1)
{Gtime=0;
for(i=0;i<Ns;i++)
{for(j=0;j<N;j++)
w1[j]=w[i][j];
FW[i]=Fitfunction(w1[],N);
if(flag==1)
{for(j=0;j<N;j++)
w[j]=w1[j];
break;}
else AllFit+=Fw[i];}
/*Selection & restructure according to the Fitness*/
if(flag!=1)
{for(i=0;i<N;i++)
Ps[i]=(1-F[i])*N/AllFit;
n=0;j=0;
while(n<Ns)
{t1=Ps[n];
t2=Ps[n];
for(i=0;i<N;i++)
```

```
w1[i]=W[n][i];
if(t1!=0)
{for(i=0;i<N;i++)
w1[j][i]=w1[i];
t1--;
j++;}
n++;}
/*crossover according to Pc*/
n=Pc*N;
i=myrand()/Ns;
j=myrand()/Ns;
Swap(w1[],i,j);
/*mutation according to Pm*/
n=Pm*Ns;
for(i=0;i<n;i++)
{t=myrand()*(i+2)%n;
Mutation(w1[Ns][N],t);}}
Gtime++;}
/*The following program shows the BP-net*/
BpTime=0
e=1;
while(e>E)
{for(i=0;i<2;i++)
x[i]=myrand()%A;
if(x[0]==x[i])y=0;
else y=1;
for(i=0;i<2;i++)
{s1[i]=x[i]*w[i];
f1[i]=Sigmoid(s1[i]);}
for(i=0,J=2;i<2,j<9;i++,j++)
{s2[i]=f1[0]*w[j]+f1[1]*w[j+1];
j++;
f2[i]=Sigmoid(s2[i]);}
for(i=0;i<4;i++)
s+=f2[i]*w[i+10];
f=Sigmoid(s);
d=f*(1-f)*(f-y);
e=fabs(f-y);
for(i=0;i<4;i++)
d2[i]=f2[i]*(1-f2[i])*w[i+10]*d;
for(i=0,j=2;i<2,j<4;i++,j++)
{t[i]=w[j]*d2[0]+w[j+2]*d2[1]+w[j+4]*d2[2]+w[j+6]*d2[3];
d1[i]=f1[i]*(1-f1[i])*t[i];}
for(i=0;i<14;i++)
```

```
{if(i<2)
w[i]-=c*d1[i]*x[i];
else
{swith(i)
{case 2:w[2]=c*d2[0]*f1[0];break;
case 3:w[3]=c*d2[0]*f1[1];break;
case 4:w[3]=c*d2[1]*f1[0];break;
case 5:w[3]=c*d2[1]*f1[1];break;
case 6:w[3]=c*d2[2]*f1[0];break;
case 7:w[3]=c*d2[2]*f1[1];break;
case 8:w[3]=c*d2[3]*f1[0];break;
case 9:w[3]=c*d2[3]*f1[1];break;
deafult:w[i]=c*d*f2[i-10];}}}
BpTime++;}
/*Test the BP_net*/
Ttime=0
while[Ttime<200]
{for(i=0;i<2;i++)
x[i]=myrand()%A;
if(x[0]==x[i])y=0;
else y=1;
test=FrontCal(x[0],x[1],W[],14);
e=fabs(test-y);
if(e<E)Rtime++;}
pf("GA-BP loop_time is%. \nThe right number is%d",BpTime,Rtime);}
```

附录 2　食物污染物分布的拟合预测模型程序

```
  aa = {{1,2},{2,15},{3,23},{4,36},{5,50},{6,59},
  {7,72},{8,91},{9,72},{10,43},{11,21},{12,6},{13,2}};
 f[x_] = Fit[aa, {1, x, x^2,x^3,x^4,x^5,x^6}, x];
 gs = ListPlot[aa, PlotStyle -> {RGBColor[1, 0, 0], PointSize[0.02]}];
 gg = Plot[f[x], {x,1,13}, PlotRange -> All];
 Show[gs, gg, PlotRange -> All]
 f[x_]
 Null
```

附录 3 食物摄入量模型相关数据

“每日摄入量样本数据”，见光盘文件：“sheet1.xls”。

附录 4 食物污染物分布模型相关数据

“偶然性检验样本数据”，见光盘文件：“sheet2.xls”；
“例行检验样本数据”，见光盘文件：“sheet3.xls”。