全国第四届研究生数学建模竞赛

题号 A

题 目 食品卫生安全保障体系数学模型的建立及其改进理论问题的研究

摘　　要：

建立食品卫生安全保障体系是关系国计民生的重大而迫切的任务。膳食暴露评估数学模型可以为公共食品卫生安全监控提供比较可靠的评估依据。论文主要研究了食品卫生安全保障体系的数学模型，包括人群食物摄入量模型、污染物分布模型和风险评估模型，阐述了各种模型的建立方法。在建模过程中，探讨解决了一些改进模型的理论问题。主要工作包括：

1、根据我国的基本国情，设计了分层多阶段的抽样调查方案，即在每个省内按照饮食文化的差异分层抽样。通过改进 CAC 食品分类标准，给出了一种较为有效的适用于抽样调查的食物分类方法。在此基础上建立了膳食模型。

2、给出了污染物分布模型及其建立方法。提出了一种有少量的样本数据迭代地外推随机变量的整体概率密度函数的算法，即通过将“未检出”数据按前次迭代的结果进行加权分配，得到新的样本分布数据，再次拟合逼近概率密度函数。讨论了由不同区域的监测数据估计全国总体分布的问题，并给出了一种解决方案。在样本分布偏离给定分布模型太大时，采用密度演化理论来估计概率密度函数。

3、建立了风险评估模型及其建立方法。探讨了模型的输入数据不配套时，抽样调查数据的衔接问题。分析了常规右分位点估计方法在本题中的局限性，在此基础上提出了一种新的污染物摄入量的 99.999%的右分位点的估计方法。

参赛密码 _____
（由组委会填写）

参赛队号 1028601　　参赛学校 东南大学

参赛队员姓名 张 亮 薛彦红 陈建军

目 录

目 录 I
一、问题的背景与分析 ..... - 1 -
二、人群食物摄入量模型的建立 ..... - 1 -
2.1 抽样调查方案的设计 ..... - 1 -
2.1.1 样本总量的确定 ..... - 1 -
2.1.2 样本数在各个地区的分布及各地区样本的抽样方法 ..... - 2 -
2.1.3 抽样数据的记录方式 ..... - 3 -
2.2 抽样调查中食物的分类方法 ..... - 3 -
2.3 人群食物摄入量模型的建立方法 ..... - 5 -
2.3.1 数据的收集和应用 ..... - 5 -
2.3.2 人群食物摄入量模型 ..... - 6 -
三、污染物分布模型的建立 ..... - 8 -
3.1 污染物分布模型的建立方法 ..... - 8 -
3.1.1 基本数据获得 ..... - 8 -
3.1.2 基本思想 ..... - 9 -
3.1.3 模型简化 ..... - 9 -
3.1.4 基于有限分布类型的拟合 ..... - 9 -
3.2 对污染物模型的进一步优化 ..... - 13 -
3.2.1 直接利用污染物大于某一数值的样本拟合 ..... - 14 -
3.2.2 利用“实际数据”拟合 ..... - 15 -
3.2.3 新型的迭代算法来拟合整体 ..... - 16 -
3.3 不同地区概率分布函数的组合问题 ..... - 19 -
3.4 对样本频率分布曲线偏离预选分布很大的处理 ..... - 19 -
3.4.1 虚拟随机过程与密度演化方法 ..... - 19 -
3.4.2 实施方法： ..... - 20 -
3.4.3 应用概率密度估计的效果图 ..... - 21 -
3.4.4 讨论 ..... - 22 -
四、风险评估模型的建立 ..... - 22 -
4.1 风险评估模型的建立方法 ..... - 22 -
4.2 污染物摄入量 99.999%的右分位点精度的提高 ..... - 24 -
五、小 结 ..... - 27 -
致 谢 ..... - 28 -
参考文献 ..... - 28 -

一、问题的背景与分析

“民以食为先，食以安为先”，食品卫生安全关系国计民生，关系党和政府的形象，关系经济繁荣和社会稳定。今年来，随着人民生活水平的不断提高和国家以人为本的发展战略的实施，食品卫生安全问题更加受到人民的关注。我国是人口众多的发展中国家，食品消费量大，而粮食生产，加工分散，因而食品卫生安全监控难度较大。学习借鉴欧美国家的食品卫生安全监控的做法，建立膳食暴露评估数学模型，用以处理调查或检测数据，对公共食品卫生安全做出评估应该是一条可行的途径。

美国和欧盟已经建立了有关模型并且开始使用，用实践证明了其可能性。但是模型的建立和使用还存在着很多严重的困难和问题。在建立人群食物摄入量模型时，需要解决的问题有抽样调查方案的设计；抽样调查中食物的分类办法和由抽样数据建立比较准确的人群食品摄入量模型等问题。污染物分布模型的建立面临的问题包括：在随机抽样数据的抽样率很低的情况下，怎样充分利用这些数据去建立模型；如何利用不详细、不完整的分类数据尽量提高模型的精度；怎么由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。风险评估模型就是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。它要解决两个模型中的调查对象和调查食品分类的不配套问题，还需要提高全体居民某项污染物摄入量的 99.999%的右分位点的估计精度。

二、人群食物摄入量模型的建立

人群食物摄入量模型（膳食模型）是用于估计不同地区、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。

2.1 抽样调查方案的设计

调查的目的是要获得我国居民的食物摄入量数据，从而为人群食物摄入量模型的建立提供样本数据。要求调查的结果能够准确反映全国的实际情况，调查结果的数据便于建模的使用，并且调查的工作量在实际可以承受的范围内。

抽样调查的方法在原文的问题中已经给出，以家庭为抽样样本单位。我们这里设计的抽样是分层三阶段随机抽样。抽样方案主要包括三部分内容：样本总量的确定；样本数在各个地区的分布及各地区样本的抽样方法；抽样数据的记录方式。

2.1.1 样本总量的确定

由抽样调查的基础理论可知，对于简单的随机抽样情况，设 p 为平均每人对某类食品的一天的摄入量 P 的估计量，调查要求的估计 P 时最大的绝对误差为 e（如 0.001），可靠度为 r（例如 99%），即：

$$p\{|\mathrm{p}-\mathrm{P}|\le \mathrm{e}\}=\mathrm{r} \qquad (2.1)$$

有抽样调查理论可知，满足上述精度要求的样本总量为：

$$\mathrm{n}_0=\frac{\mathrm{u}_r^2\mathrm{P}(1-\mathrm{P})}{\mathrm{e}^2} \qquad (2.2)$$

其中：n_0位样本单元总数；$r'=1-r$；u_r^2是标准正态分布的双侧$1-r$分位点。

本次调查的抽样方法是分层三阶段无放回随机抽样，属于复杂抽样设计。为了给出复杂抽样设计下的样本家庭数目总量，需要借助于设计效应 deff。在相同精度下，复杂抽样设计的样本量是简单随机抽样设计的样本量的 deff 倍。根据理论分析和实际经验，经过仔细分层的三阶段不等概率抽样的设计效应大约为 2。因此在复杂抽样方法下，所需样本户总量为：

$$n=deff\cdot n_0 \qquad (2.3)$$

具体到我们这里的抽样问题，需要对公式（2.2）进行一定的修改，且需要同假设的样本的模型相结合。具体地说，由于人均某类食物的摄入量不是满足严格的标准正态模型，因此要对公式（2.2）中的u_r^2修正为满足估计随机变量的随机概率密度函数的双侧$1-r$分位点。另外，考虑到对人均食物摄入量调查的复杂程度和需要的工作量，我们要限制总体的抽样数目在几千户，至多几万户。所以需要在样本总量，最大绝对误差和可靠度等统计量之间进行折中，以把抽样样本总量控制在可以接受的范围之内。

2.1.2 样本数在各个地区的分布及各地区样本的抽样方法

考虑到我国地域广阔，各地区之间的饮食差异也比较大，所以样本数目在各个地区之间的分布应该详细地考虑。这里给出三阶段的随机抽样设计方案，其中第一阶段是省级行政单位的抽样，第二阶段是对地级市的抽样，第三阶段是在地级市中抽取家庭。抽样理论证明，多阶段抽样中对总的抽样误差起主导作用的是前一、二阶段抽样，随着抽样阶段的递增，以后各阶段产生的抽样误差在总的抽样误差中所占的份额越来越小。因此在多阶段抽样中抽样阶段越往前抽样比应越大，抽样阶段越往后抽样比应越小。另外，我国大部分省级行政单位之间都会有饮食差别，而且每个省级行政单位都有一定的财力物力来开展调查活动。因此我们设计对所有的省级行政单位都开展抽样调查，各个省的抽样样本数与各省的人口数在全国总人口数所占的比例相同。即假设第 i 个省的总人数为 M_i，全国的总人数为 M，则该省的抽样样本数为 M_iW_i，其中 $W_i=\frac{M_i}{M}$ 称为该省的权。由于所有的省都进行调查，所以在第一层（省级行政单位层）不需要进行抽样。对于地级市的抽样我们是这样考虑的，对于我国的绝大多数省来说，省内的饮食习惯还是存在一定差异的，但每个省可以按地理位置分为几块区域。各地理区域内人们在饮食习惯，农作物生产结构，经济状况等方面的差异是比较小的。我们以山东省为例，山东现辖 17 个地级市，根据地理位置可以分为胶东（烟台，威海，青岛，日照）,西北（德州，聊城，菏泽），鲁中（济南，泰安，莱芜，淄博），鲁北（滨州，东营，维坊）及鲁南（济宁，枣庄，临沂）等五块区域。由于前述的原因，我们有理由认为这五块区域内的地级市的膳食模型是相同的，因此我们可以从每块中只抽取出一个市来调查。因此我们对每个省的地级市的抽样方法是：把各省按人文地理分成几个大的区域，每个区域包含几个地级市，然后用随机抽取的方式从每个区域中选取一个地级市来做调查（分层按比例抽样）。每个地级市的抽样样本数与该市所在的区域内的总人数与全省的总人数的比值成比例。即由前面第 i 个省的总人数为 M_i，而该省总的抽样样本数为 M_iW_i，假设抽到的市在第 j 个区域内，而该区域的总人数为 M_{ij}，则这个市的样本总数为 $M_iW_iW_{ij}$，其

中，$W_{ij}=\frac{M_{ij}}{M_i}$，为该市的层权。第三个阶段是要从地级市中抽取家庭样本，通过前面的分析可看出，我们在每个省内选择的地级市数量比较少（但是仍然保证了抽样能反映实际的情况），因此可以保证每个市有一定样本数，即能保证样本的规模，便于进行分布估计。这样在地级市内抽取家庭样本时按照均匀抽取，就可以保证抽取样本在年龄、男女、城乡等方面的能较好的代表全市的实际情况。

关于调查时间的安排：由于食品，特别是蔬菜类有很强的季节性，所以最好每季进行一次抽样调查，如人力物力受限，可夏秋和冬春各一次，每次 3～7 天时间。

2.1.3 抽样数据的记录方式

抽样调查的方法已经在问题中给出，这里我们只是说明调查结果的记录方法。调查结果的记录方法必须与后面的模型建立方法，以及对数据的分析过程结合起来考虑。其中比较重要的问题包括：调查时食物的分类方法，这部分内容我们将在下节中单独进行讨论；被调查对象的分布情况。由于不同人群的食物摄入量与其性别、年龄、劳动强度、经济收入等方面密切相关，因此调查人员在调查每个家庭时需要对家庭成员的基本情况作比较详细地记录，其中包括被调查人员的性别、年龄、从事工作的类型、经济收入、各成员所摄入的各种食品的大体数量等。由此我们即可得到按照不同的分类情况统计的人群食物摄入量模型。另外，人们所消费的食物有明显的季节性，因此抽样时也要记录抽样时的季节，以便于进行数据统计。

2.2 抽样调查中食物的分类方法

我国居民消费的食物种类非常复杂，包括：主食、肉类、蔬菜、水果、水、饮料、各种调味剂和经过加工的食品，细分将达数千种以上。在实际调查过程中如果详细地分类，其调查工作量太大，得到的数据也会难以处理，不利于描述食物摄入量的分布情况；而如果随意粗糙进行分类，则将影响调查的精度，因此需要根据污染物分布模型的数据合理设计抽样调查中食物的分类办法。

目前常用的分类方法主要有以下几种。第一种是从营养学角度来分，可以将食物分为以下五类：1、谷类及薯类：谷类包括米、面、杂粮，薯类包括马铃薯、红薯等；2、动物性食物，包括肉、禽、鱼、奶、蛋等；3、豆类及其制品，包括大豆及其他干豆类；4、蔬菜水果类，包括鲜豆、根茎、叶菜、茄果等；5、纯热能食物，包括动植物油、淀粉、食用糖和酒类。

2002 年，国家卫生部、科技部和国家统计局组织的“中国居民营养与健康状况调查”。调查采用连续 3 天 24 小时回顾询问法来调查居民所有摄入食物，用“称重法”调查家庭肉及肉制品消费量。其中对食物的分类包括：粮谷类（米、面和其它）、薯、豆类、干豆类、豆制品、蔬菜（深色蔬菜、浅色蔬菜）、腌菜、水果、坚果、畜禽肉蛋及水产品、猪肉其他畜肉、动物内脏、禽肉、蛋及其制品、水产品、奶及其制品、食用油（植物油、动物油）糕点类、糖及淀粉、食盐、酱油、酱类等。

第三种食物分类方法是国际食品法典委员会(CAC) 食品与饲料分类标准。早在国际食品法典委员会（CAC）开始农药残留限量工作制定的同时，就开展了食品和动物饲料产品分类的工作。1989年，第18次法典大会采用了这一分类系统，将它作为《食品中的农药残留建议指南》的第4部分，意味着食品和动物饲料产

品分类已成为国际标准的一个独立部分。目前，CAC 已将整个食品和动物饲料产品分类系统纳入了计算机管理系统，这样所有农药残留方面的法典都可以通过程序链接到该系统。现在至少有4000种食品和饲料产品已包含在该数据库中。

CAC食品和饲料分类标准的框架：食品法典分类标准是按4个层次来进行食品和动物饲料分类的。首先，按照原料来源及是否经过加工分为A、B、C、D、E 5个等级。这5个等级分别是A级 植物来源的初级食品（包括5种类型）、B级 动物来源的初级食品（包括5种类型）、C级 初级饲料（包括1种类型）、D级 植物来源的加工食品（包括4种类型）、E级 动物来源的加工食品（包括4种类型）。其次，在等级划分的基础上，根据产品特征共划分为19个类型。各大类中根据农药残留相似性再分为不同的组，如水稻、麦类、杂谷类、豆类、薯类等，分类标准的每个组中都包含各种推荐指标，如潜在农药残留量、作物和动物在正常情况下的消费量以及应用农药最大残留限量的商品比例等。组是由不同的商品名称组成，食品和饲料法典分类标准对商品特征描述比较清楚，包括数字编号、字母代码、产品名称、拉丁名称、产品特征等。数字编码包含的信息包括：等级、类型、分组、字母代码、数字代码和特征描述等。标准中所列举的商品比较全面，包括各个国家的各种相关产品，产地不同的同一品种的商品也都分别列举了出来。

通过以上的描述可以看出三种食物分类方法中，CAC 食品和饲料分类标准较多地考虑了食品中的农药残留情况，这正好符合食品安全评估的要求。另外，CAC 食品和饲料分类法是按照四个层次来进行分类的，分别是等级（5 个）、类型（19 个）、组和商品名称。这种分类方式有利于不同时期，不同地域，不同种类的抽样数据之间，食物摄入量模型与污染物分布模型之间数据的相互转换和组合，所以在污染物分布模型的数据分类不一致时，膳食模型的数据可以方便地与其匹配。因此这种分类方式具有非常较强的优势，我们在此建议采用 CAC 食品和饲料分类标准对食物进行分类。但是，针对这里的具体的实际问题，可以在两个方面对 CAC 的分类标准进行改进。首先 CAC 的分类标准是针对食品和饲料的，而我们要评估的是食品安全问题，所以可以将其中对饲料的分类去除。具体地说就是可以只考虑其 5 个等级中的 4 个（A、B、D、E），而无需考虑第 C 个等级。其次，CAC 分类方法的分类标准是农药残留，由于现实生活中的食品污染的多样性，我们认为有必要对其它的污染源加以考虑，例如钢铁、水泥、电力、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等来对食品进行分类。特别是应针对题目所给出的危害面广、后果严重的几种污染物，如：铅、镉、有机磷、有机氯等来分类，将更加有利于人群食物摄入量模型与污染物分布模型之间的匹配。因此我们这里给出的食品的分类方法的总体框图表示在图 2.1 中。

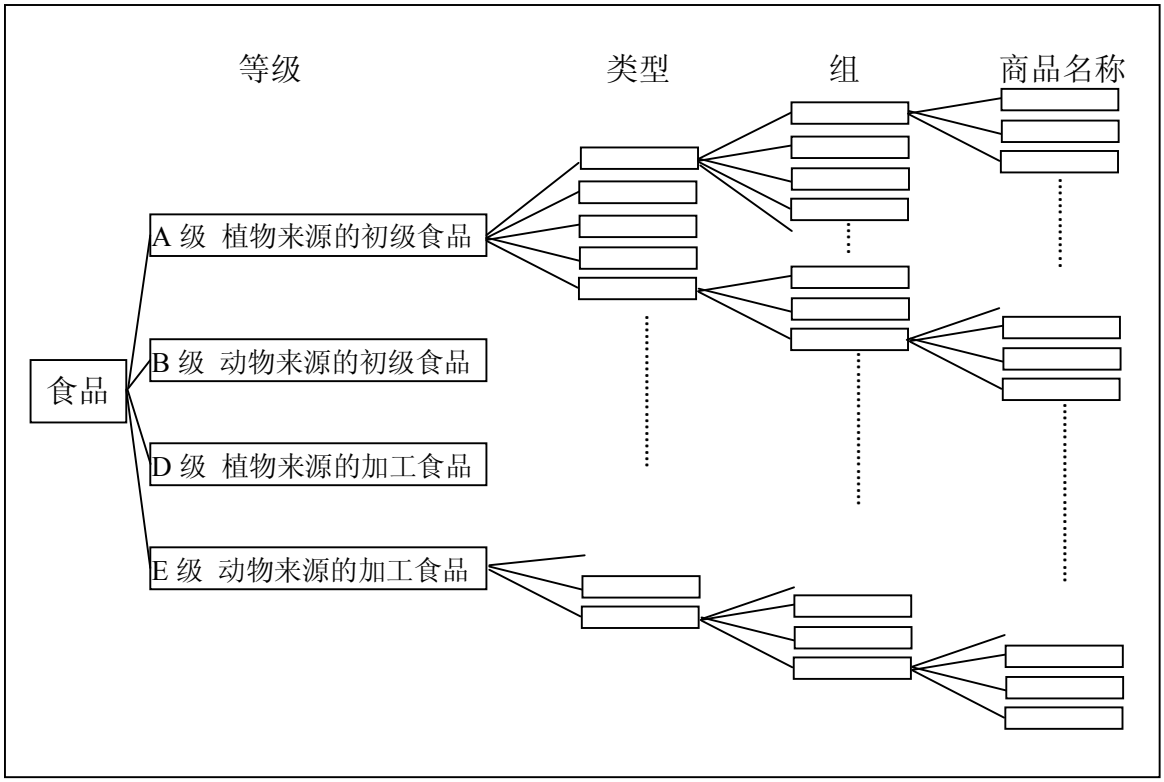


图 2.1 食物分类方法框图

2.3 人群食物摄入量模型的建立方法

人群食物摄入量模型（膳食模型）是用于估计不同地区、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。很明显，食物摄入量与前面这几个因素有关系，因此得到人群食物摄入量模型应该看成是一个随机矢量，它与前面所列的因子有关，可以表示为 intake(region, sex, age, season, intension, income)。在建立人群食物摄入量模型时，我们考虑对这些方面进行改进：建立模型时多种数据的收集、选择、融合和应用；人群食物摄入量模型。

2.3.1 数据的收集和应用

在构建膳食模型时，除了我们抽样调查所得到的数据外，我们还要充分利用一切从其它渠道可以获得的信息，我们调查的样本单元是单个家庭，调查其食物摄入量，并对家庭成员的基本情况作以记录，来分析各种情况的人群的食物摄入量。但是实际上还有很多的数据我们都可以利用，例如，全国或各个地区的粮食、蔬菜、水等食物的总产量或者总消费量这些数据都可能有的，而且我们肯定有比较详细的各地区，全国的人口普查的资料可以用。不利用这些资源也是一种资源的浪费。由上面的这些数据，我们就可以得到人群的各类食品的一天的摄入量的估计，这种估计食物消费量的方法称为供应与消费统计法。供应与消费统计的消费量的一般计算公式为:

人均膳食量(g/d)=总消费量/总人口数/365天

其中：总消费量=国内产量+进口量-出口量-工业用量和其他消耗量

该法的优点是从宏观上得到确切的数据，避免了调查的代表性问题；缺点是资料来源的准确性难以保障。例如，进出口数量、工业消耗量、其它损耗量等都

较难获得准确数据。

在用供应与消费统计法估计出人均膳食量之后，我们就可以按照加权组合等方法就可以得到食品摄入量的更为准确的估计。

2.3.2 人群食物摄入量模型

由于人群食物摄入量模型是对不同情况人群的各种食品的日均摄入量，因此我们的模型要考虑人群的不同情况，以及各种食品的情况。我们的建模方法是：先分别针对某种食品建模；而全部种类食品的摄入量模型则通过将各种食品按照其总量之间的比例关系来加权，再求其和来得到。而对每种食品的建模又需要按人群的不同情况来分别建模。具体地说就是，按性别（男、女），年龄（分成几个不同的年龄段），城乡（城市、农村），地区（先对各个地级市分别建模，再按其相应的人口总数的比例进行加权生成全省的模型；然后将各个省的模型按其总人口比例加权相加来组成全国的模型）。这种建模方法的优点是由于我们掌握了各个单独分量的摄入量模型（比如各个地区，各个年龄段或性别情况），因此当我们需要评估某个地区或者某种特征的人群的模型时，可以直接利用其相应的摄入量模型，而不需要用全国的模型来近似，精度可以做到更高。另外一个优点是我们按照各子集类摄入量模型相应样本数之间的比例来加权求和，可以保证最后所得到的估计量是总体目标量的自加权无偏估计。

而对于某种具体的食品而言，国家卫生部、科技部和国家统计局在 2002 年组织了“中国居民营养与健康状况调查”。调查队在全国 31 个省、直辖市、自治区的 132 个调查点中选取 23470 户进行膳食调查，共调查 68962 人。膳食调查采用连续 3 天 24 小时回顾询问法调查居民所有摄入食物，及用“称重法”调查家庭肉及肉制品消费量。根据调查结果，部分种类食品的人日均膳食摄入量见表 2.1。对所调查食品日均膳食量的统计结果显示，绝大多数食品的日均膳食量摄入水平服从正态分布，其中我国居民日均摄入熟肉制品的量的分布情况见图 2.2。

表 2.1 二 00 二年全国营养调查部分食品人均膳食摄入量数据

食品类别	每日消费量（g）	标准差	消费人群比例
蔬菜	280.9259699	168.8339531	0.99
水产品	87.91046272	70.91503249	0.34
生猪肉	82.33422109	65.46614034	0.65
生禽肉	66.22148968	53.72746529	0.18
生羊肉	52.50807958	42.42764754	0.038
熟禽肉制品	50.41941748	38.8737811	0.016
熟畜肉制品	47.23153693	42.26154557	0.099
生牛肉	47.03479886	41.88530612	0.086

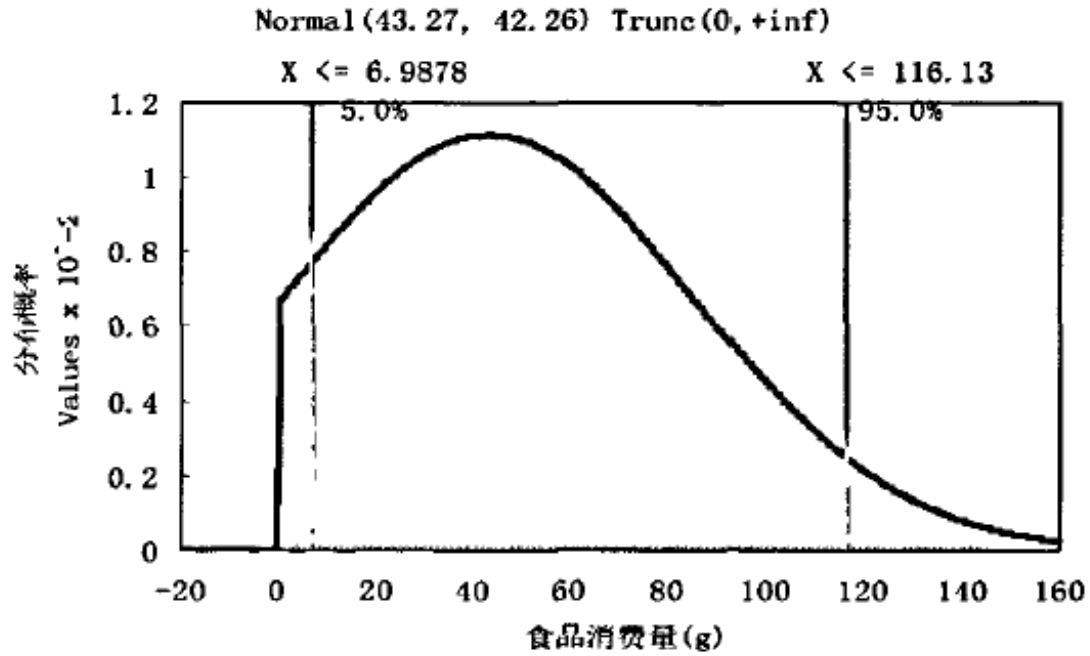


图 2.2 熟肉制品人日均摄入量分布图

我们这里考虑由各个省的调查数据来估计全国的摄入量模型的方法，假设第 i 个省的摄入量模型为：

$$f_i(x)=\begin{cases}\dfrac{1}{\sqrt{2\pi}\sigma S}\exp\left\{-\dfrac{1}{2\sigma^2}\left(x-u\right)^2\right\}, & x\geq 0\\ 0 & x<0\end{cases} \tag{2.4}$$

其中，随机变量 x 为某类食品的日均摄入量，u 为该类食品日均摄入量的平均值，σ 为日均摄入量的方差。S 是将该不完整的概论分布归一化的常数，它等于该分布在所有随机变量范围上的面积：

$$S=\frac{1}{\sqrt{2\pi}\sigma}\int_0^{+\infty}\exp\{-\frac{1}{2\sigma^2}(x-u)^2\}dx \tag{2.5}$$

则该食品全国的摄入量模型通过下式计算：

$$f(x)=\sum_{i=1}^{N}\mathrm{W}_i f_i(x) \tag{2.6}$$

其中：$\mathrm{W}_i=\dfrac{\mathrm{M}_i}{\mathrm{M}}$ 为该省的总人口数 M_i 与全国的总人口数 M 之间的比值，即为该省的权；N 为全国的省级行政单位的数目。

以下两个定理保证了我们所提的建模方法的有效性。

定理 2.1 在分层抽样中，样本的加权平均数：

$$\hat{Y}=\sum_{i=1}^{N}W_i\overline{y}_i \tag{2.7}$$

是总体样本平均数 $\overline{Y}$ 的无偏估计量，即有：

$$E\hat{Y}=\sum_{i=1}^{N}W_iE\overline{y}_i=\overline{Y} \tag{2.8}$$

式中：样本总体 Y 被分成了 N 层：$y_1, y_2,\ldots, y_N$；$\overline{y}_i$ 表示样本 y_i 的平均估计；W_i 表示该子样本数在全部样本数中所占的比例，即其权值。

定理 2.2 统计量 $\hat{Y}$ 的方差为；

$$V_{\hat{Y}}=\sum_{i=1}^{N}W_iV_{\overline{y}_i}=\sum_{i=1}^{N}W_i\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2 \tag{2.9}$$

其中：$V_{\hat{Y}}$ 表示 $\hat{Y}$ 的方差；$V_{\overline{y}_i}$ 表示 $\overline{y}_i$ 的方差；S_i 表示第 i 层中简单样本的标准差。且其无偏估计量可取为：

$$S^2(\hat{Y}):=\sum_{i=1}^{N}W_i\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2 \tag{2.10}$$

从定理 2.1 可以看出，我们这里给出的分层抽样估计法对样本均值的估计是无偏的；由定理 2.2 可以看出，我们可以用其构造统计量方差的无偏估计。

三、污染物分布模型的建立

污染物分布模型是根据农药、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。

3.1 污染物分布模型的建立方法

3.1.1 基本数据获得

污染物分布模型的数据主要是来自：食品卫生监测部门日常对市场上食物的检测数据（包括例行监测数据和偶然抽查数据，符合性检验和监测性检验数据）和市场上各类食品的流通量，进出口口岸的检测数据。此外，历史检测数据也是非常重要的信息。

为了和食物摄入量模型中的数据相匹配，污染物的采集对象应该与食品的分类尽量一致，如我们食品分类有大豆类，那我们检测大豆的含量时，尽量不要把大豆与别的作物分为一个合集。

为了便于说明，我们使用汞污染的数据，数据来自于武汉市的调查数据。

表 3.1 汞污染的分布数据

Hg		分布类型	偏度系数	峰度系数	最小值 ($mg \cdot kg^{-1}$)	最大值 ($mg \cdot kg^{-1}$)	均值±标准差 ($mg \cdot kg^{-1}$)	二级标准 ($mg \cdot kg^{-1}$)	样品数
污染区	表层	lgN	-1.020	0.800	0.010	1.126	0.314±0.322	1.000	15
	底层	lgN	0.980	0.660	0.006	0.477	0.181±0.173	1.000	15
非污染区		N	-0.722	-0.447	0.030	0.064	0.050±0.013	1.000	8

注：N 为正态分布，lgN 为对数正态分布（偏度系数和峰度系数为对数转换后的值）.

3.1.2 基本思想

基本思想是首先假定一个分布类型，进行该分布类型情况下的统计量分析，并采用某种准则进行拟合优度检验。然后，我们考虑了对“未检出”食品的处理，最后，有限的规则分布类型难以满足实际应用的需要，为此，我们进行了统计分布与统计方法的改进。

3.1.3 模型简化

污染物分布的特点有：第一、污染物含量不会是负的，所以概率密度函数分布在右半平面；第二、污染物分布是左偏态，即均值左边的概率远大于均值右边的概率；第三、食品绝大多数都是污染物含量很低的，所以概率密度函数集中分布在污染物数值小的区域；第四、对于某种待测食品，要检测的污染物数值为零的情况也很少，所以概率密度函数的峰值不在随机变量取零时。

3.1.4 基于有限分布类型的拟合

首先，我们先获得样本频率分布曲线,对连续型随机变量x，观测n次，把样本的取值区间划分成若干等间距小区间$(x_1,x_2]$,$(x_2.x_3]$.....，样本值落入$(x_k,x_{k+l}]$的个数为n_k,则样本频率为:

$$f(x_k)=\frac{n_k}{n}\frac{1}{\Delta x} \tag{3.1}$$

其中$\Delta x=x_{k+1}-x_k$，$f(x_k)$曲线为样本频率分布曲线，简称频率曲线。当$n\to\infty,\Delta x\to 0$时，$f(x_k)$趋向于概率密度。

采用最大似然估计来拟合数据：

设总体X的密度函数$f(x,\theta),\theta$是参数或参数向量，$X_1,X_2,\cdots,X_n$是该总体的样本，对给定的一组观测值$x_1,x_2,\cdots,x_n$，其联合密度是θ的函数，又称似然函数，记为：

$$L(\theta)=L(\theta,x_1,\cdots,x_n)=\prod_{k=1}^{n}f(x_k,\theta),\theta\in\Theta \tag{3.2}$$

其中Θ为参数集，若存在$\hat{\theta}=\hat{\theta}(x_1,\cdots,x_n)\in\Theta$使$L(\hat{\theta})\geq L(\theta),\theta\in\Theta$就称$\hat{\theta}(x_1,\cdots,x_n)$是$\theta$的最大似然估计值，而$\hat{\theta}(X_1,\cdots,X_n)$是$\theta$的最大似然估计量。

对给定的观测值，$L(\theta)$是θ的函数，最大似然估计的原理是选择使观测值$x_1,x_2,\cdots,x_n$出现的“概率”达到最大的$\hat{\theta}$作为θ的估计。

最大似然估计具有不变性，即若$\hat{\theta}$是θ的最大似然估计，则$g(\theta)$的最大似然估计为$g(\hat{\theta})$。但是矩估计不具有不变性，例如假定$\overline{X}$是θ的矩估计，一般情形下，θ^2的矩估计不是$\overline{X}^2$。

假设总体X具有

$$f(x,\theta)=\frac{\beta^k}{(k-1)!}x^{k-1}e^{-\beta x},x>0,\beta>0 \tag{3.3}$$

其中 k 是已知的正整数， 未知参数 β 的最大似然估计按以下步骤进行：

Step1： 对给定的观测值，其似然函数为： $L(\beta)=\prod_{j=1}^{n}\frac{\beta^k}{(k-1)!}x_j^{k-1}e^{-\beta x_j}$

Step2： 当 $x_j>0$ 时，对数似函数为：

$$\ln L(\beta)=\sum_{j=1}^{n}(k\ln\beta-\beta x_j+c^*)=nk\ln\beta-\beta\sum_{j=1}^{n}x_j+c \tag{3.4}$$

Step3：令 $\frac{\partial}{\partial\beta}\ln L(\beta)=\frac{\partial}{\partial\beta}(nk\ln\beta-\beta\sum_{j=1}^{n}x_j+c)=\frac{nk}{\beta}-\sum_{j=1}^{n}x_j=0$

由此得到：

$$\hat{\beta}=\frac{nk}{\sum_{j=1}^{n}x_j}=\frac{k}{\overline{x}} \tag{3.5}$$

所以 β 的最大似然估计量为

$$\hat{\beta}=\frac{k}{\overline{X}} \tag{3.6}$$

基于污染物分布的特点，我们提出六类污染物分布模型： χ^2 分布（chi-square）、F 分布、Gamma 分布、对数正态分布、Weibull 分布、Nakagami 分布（瑞利分布是 Nakagami 分布在 m=1 时的特例）

由于污染物种类繁多，若用一种分布来拟合，则误差很大，对于某一污染物，我们采用多种分布模型进行拟合，这样就提高了分布的精度。

下面我们以汞污染分布为例，详细地描述建模过程。

对数据的分析：我们把汞污染的分布用概率直方图来表示，如图 3.1 所示：

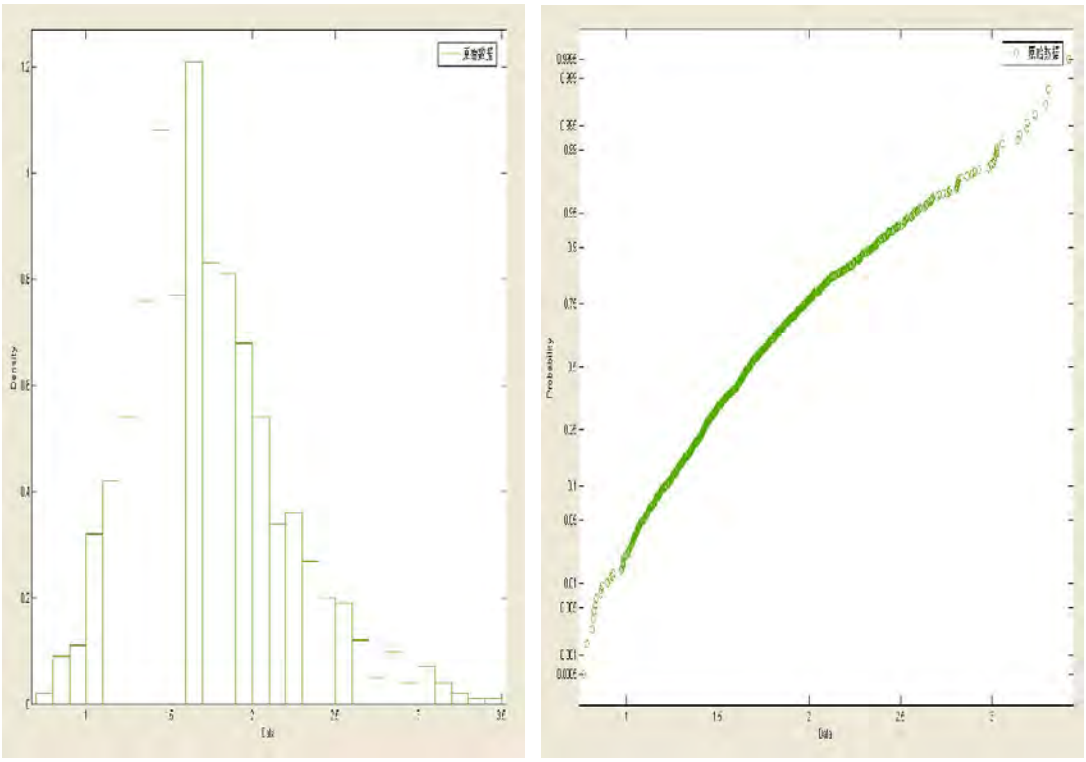


图 3.1 原始数据

我们选择 Nakagami 分布拟合得到，如图 3.2 中的红线所示：

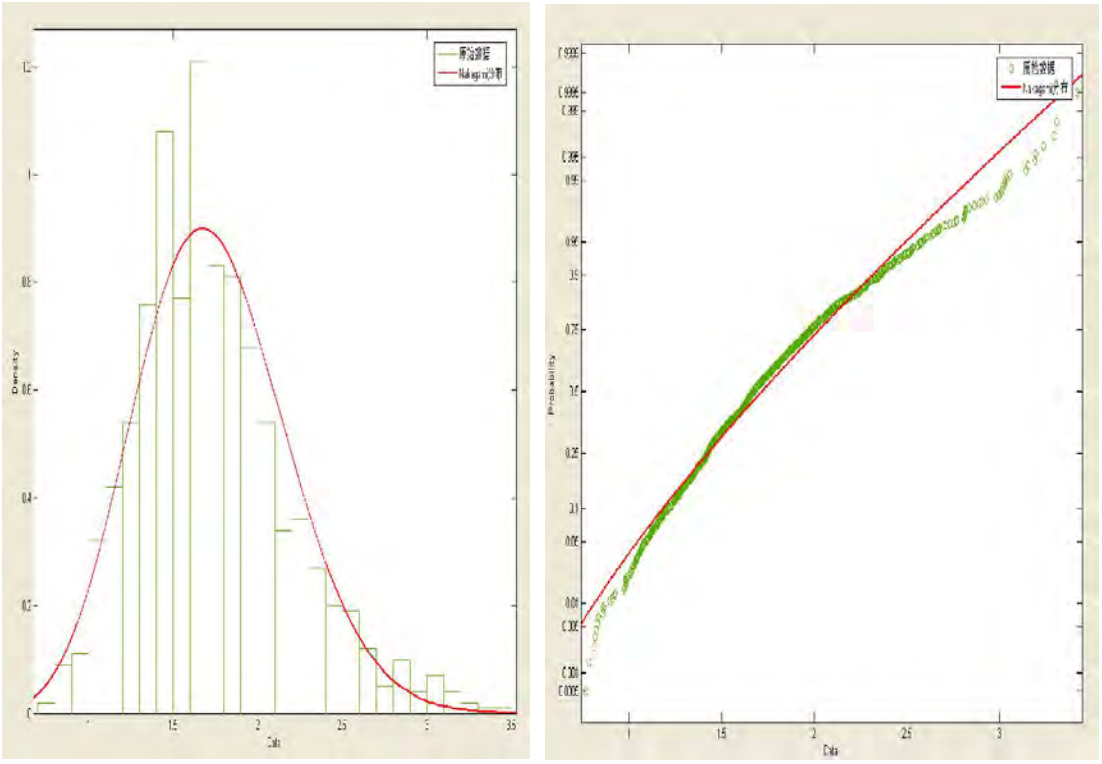


图 3.2 Nakagami 分布拟合结果

选择对数正态分布拟合，如图 3.3 中的蓝线所示。

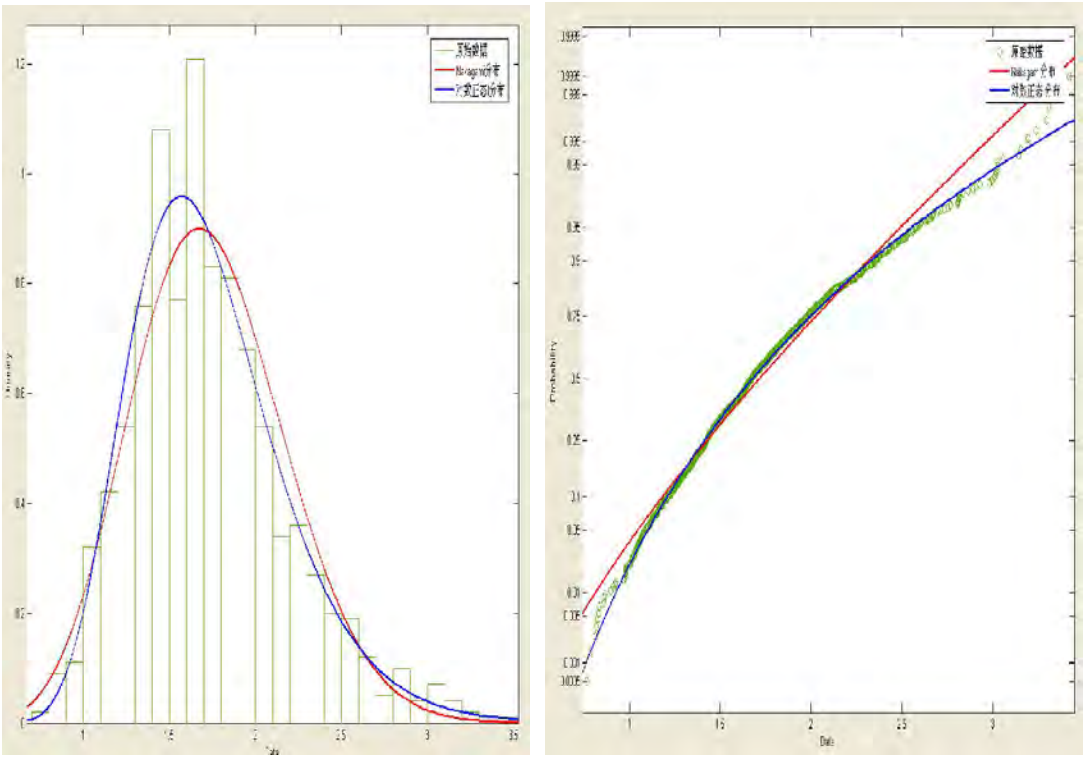


图 3.3 对数正态分布拟合结果

选择 Weibull 分布拟合，如图 3.4 中的棕线所示。

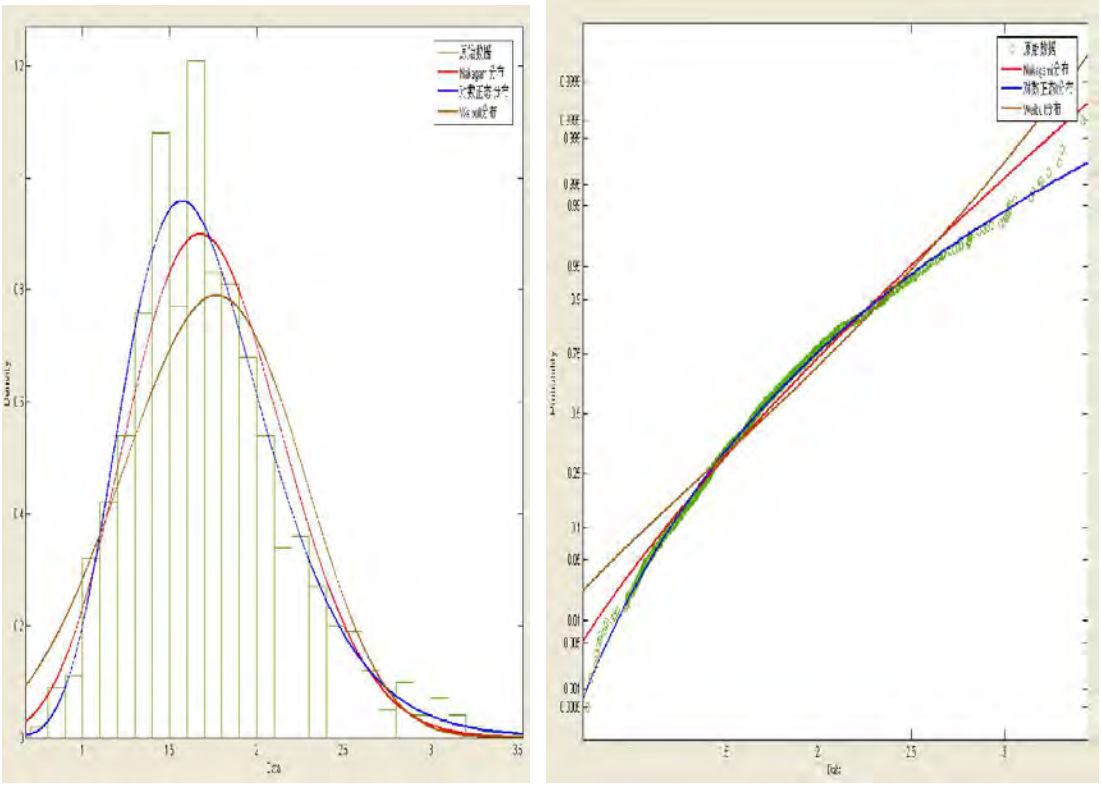


图 3.4 Weibull 分布拟合结果

采用 Gamma 分布拟合，如图 3.5 中的灰线所示

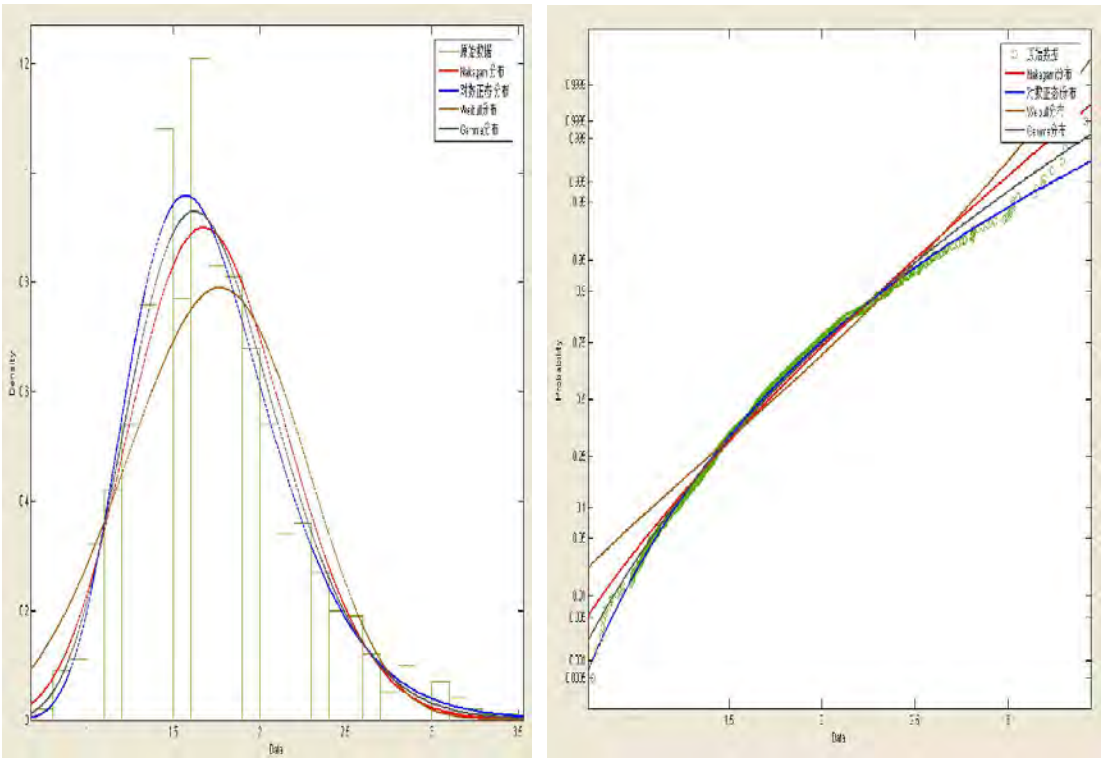


图 3.5 Gamma 分布拟合结果

仿真结果的性能统计如下表所示：

表 3.2 不同分布的拟合结果统计表

Distribution	Log likelihood	Domain	Mean	Variance
Nakagami	-592.768	0 < y < Inf	1.73503	0.194237
Lognormal	-573.739	0 < y < Inf	1.73295	0.203774
Weibull	-648.255	0 < y < Inf	1.72663	0.239283
Gamma	-578.096	0 < y < Inf	1.73277	0.194404

由上表我们可以明显看出 Lognormal 的 Log likelihood 是最优的，所以我们这次采样数据用 Lognormal 分布拟合，误差较小。

由于污染物种类多种多样，所以拟合的的分布和参数也各不相同。我们采用了一类分布（χ^2 分布（chi-square）、F 分布、Gamma 分布、对数正态分布、Weibull 分布、Nakagami 分布），来拟合不同的数据，提高了拟合的精度。

3.2 对污染物模型的进一步优化

偶然抽查数据和监测性检验数据有所有被检测食品的污染物含量，而符合性检验只有“被检出”的食品的污染物含量，没有“未检出”食品的污染物含量，如果把“未检出”食品的污染物含量作为零，那么我们得到的污染物分布模型会有很大误差。

对于检测结果来说，因为对“已检出”的食品进行了全部检测，所以样本频率分布函数的右边比较精确，但是在食品安全的情况下"未检出"的数据远比有具体

数值的次数多,或者说数据经过筛选,左边数据不及右边数据精确,左边数据不及右边数据精确。

当我们利用实际的为了充分利用“未检出”食品数据，我们采用将利用 2%的随机数据，以及“未检出”食品数据对左边的数据进行加权处理，经过数次迭代逼近理论分布。

对存在“未检出”食品数据进行假设：第一、认为符合性检验的中“已检出”数据的污染物含量是真实的，而“未检出”食品数据的污染物含量认为为零。第二、偶然检测数据是针对符合性检测中未检测部分进行的，对。第三、虽然偶然抽查数据是总体数据量的 2%，但对实际抽样方案来说，由于总体的数据量很大，则偶然检测数据量应该不会太少。我们把满足上述假设的数据记为“实际数据”。

3.2.1 直接利用污染物大于某一数值的样本拟合

如果只是利用随机变量取值大于某一数值 t 的部分样本数据来估计整体的分布则性能不好，如下图所示：

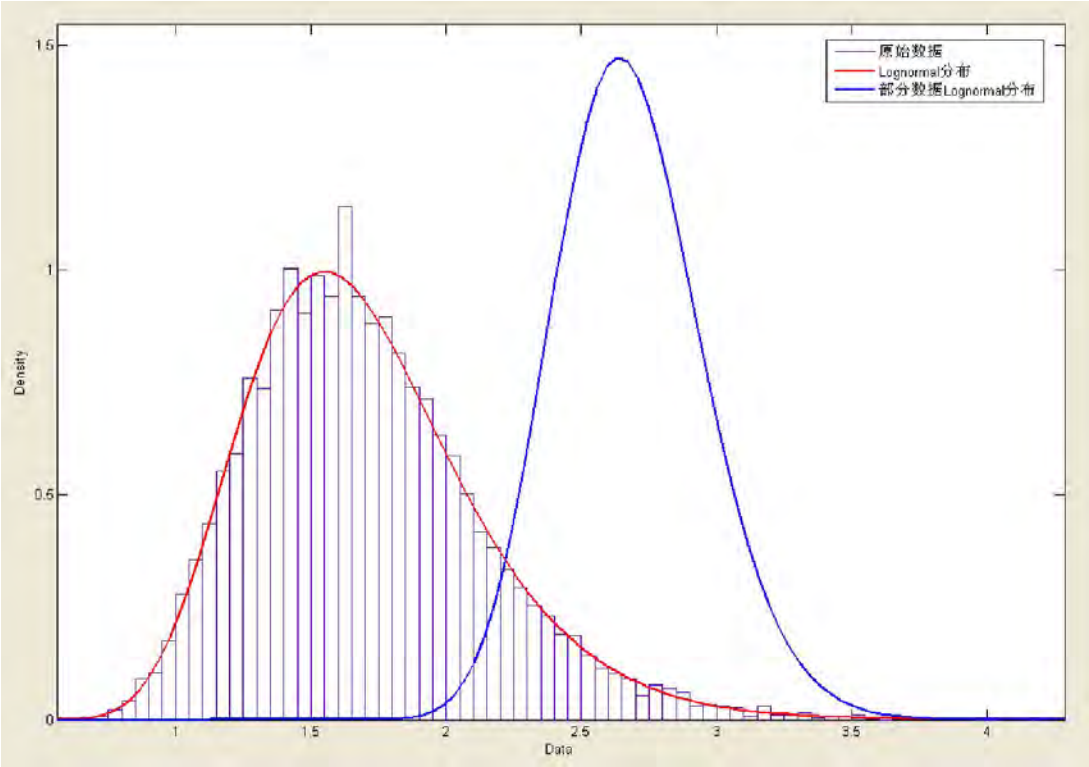


图 3.6 用部分样本数据来估计整体的分布的效果

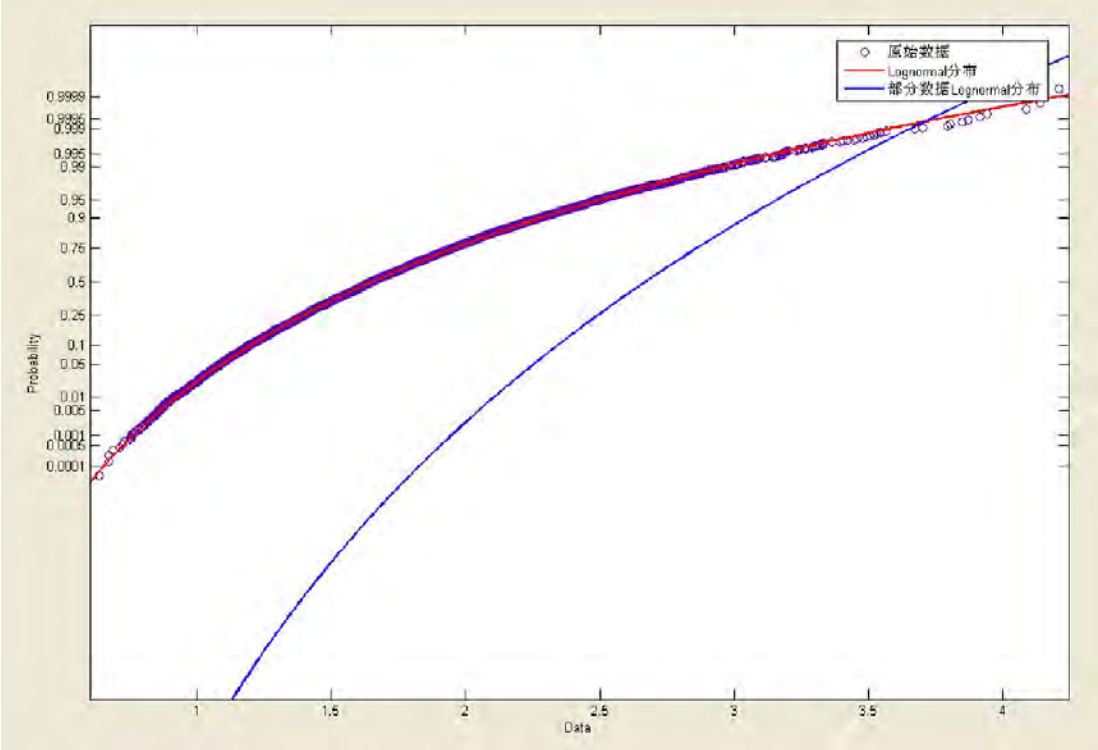
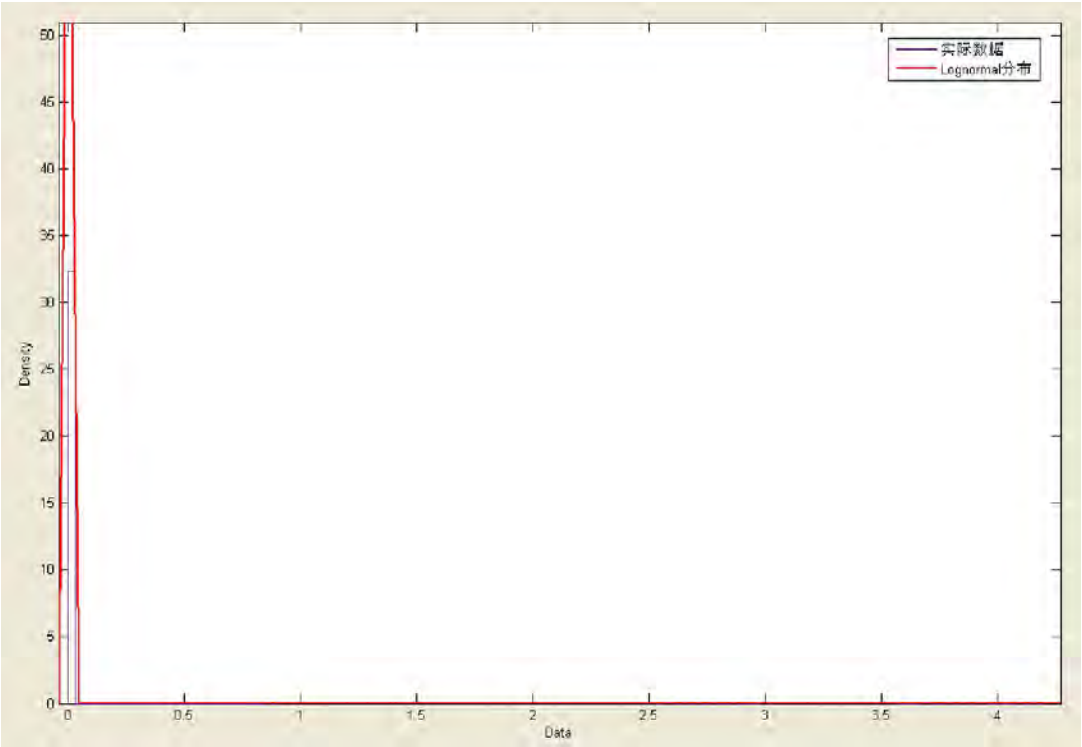


图 3.7 用 Lognormal 分布分别拟合原始数据和部分数据的结果

上图中，红线是用 Lognormal 分布来拟合原始数据数据，而蓝线是用 Lognormal 分布来拟合后 7%数据（实际情况下可能比 7%更小），由此可以清楚的看到误差很大。

3.2.2 利用“实际数据”拟合

如果直接利用满足假设的“实际数据”来估计整体的分布，误差也很大，如下图所示



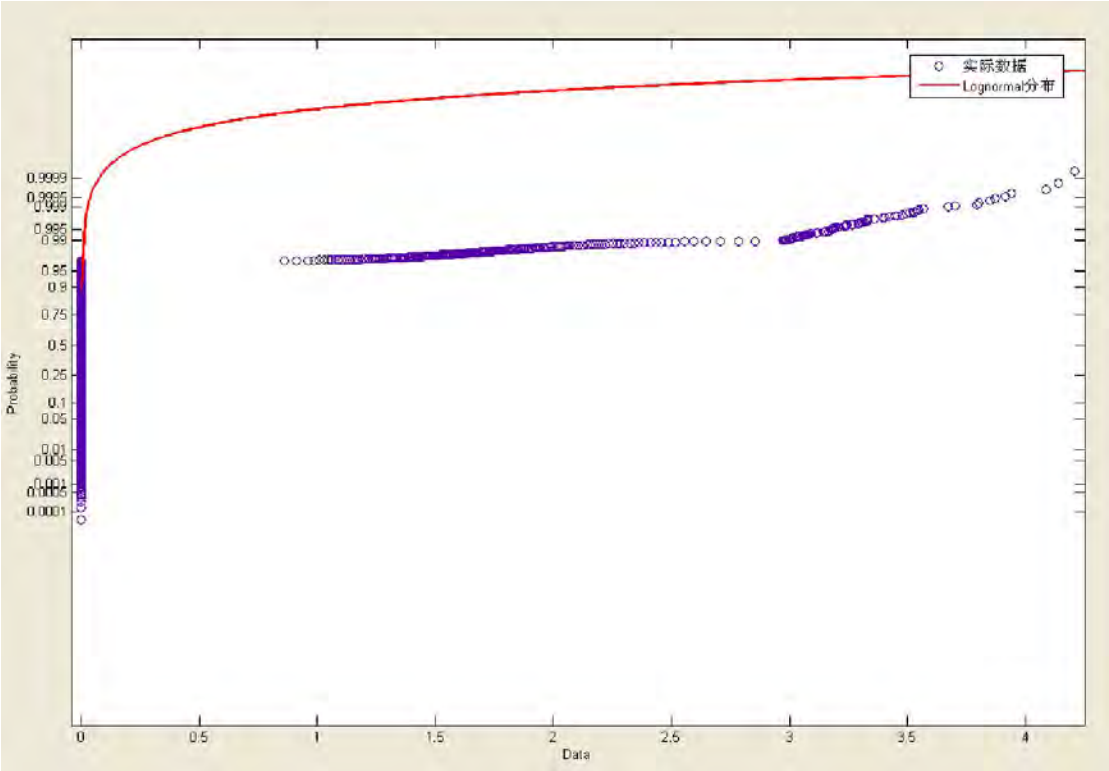


图 3.8 直接利用满足假设的“实际数据”来估计整体的分布

由于把满足符合性检测的大部分数据的污染物含量当作零，所以拟合概率密度函数在 0 处有个峰值，这和实际情况并不相符。

3.2.3 新型的迭代算法来拟合整体

我们提出一种新的迭代算法来拟合整体概率密度函数的分布。

算法描述：

第一步：把随机变量在零处的概率按比例分配到 2%的偶然监测数据中，进行拟合，得到概率密度函数 $f_0(x)$，这个比例是 2%的偶然检测数据的分布，即：

实际数据中：$P_1(x=0)$，$P_1(x=x_1)$

重新分配数据后：$P_2(x= x_1) = P_1(x=x_1) + aP_1(x=0)$，其中 a 为加权系数

第二步：把随机变量在零处的概率按新的比例分配到 2%的偶然监测数据中,进行拟合，得到概率密度函数 $f_1(x)$，这个新的比例是按照 $f_0(x)$在小于 t 部分的函数决定的。

第三步：比较 $f_1(x)$与 $f_0(x)$的距离，若距离小于阈值，则结束迭代，否则到第二步继续迭代。

我们对该算法进行了仿真实验：仿真条件：理想数据服从 Lognormal 分布，“已检出”数据的污染物含量位于 99%右分位点右边，偶然抽查数据占总样本数的 2%。

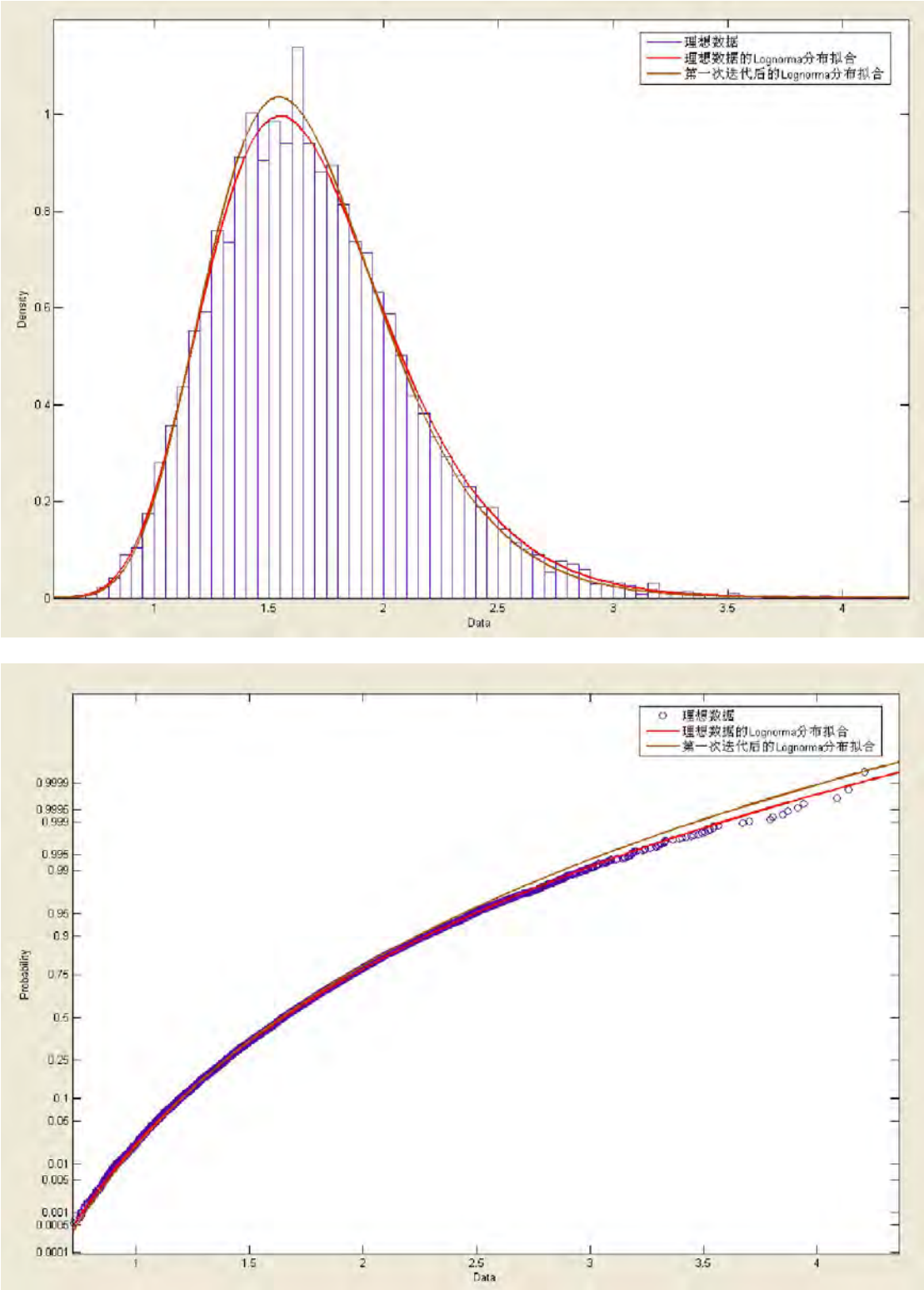


图 3.9 用 Lognormal 分布拟合时第一次迭代的结果

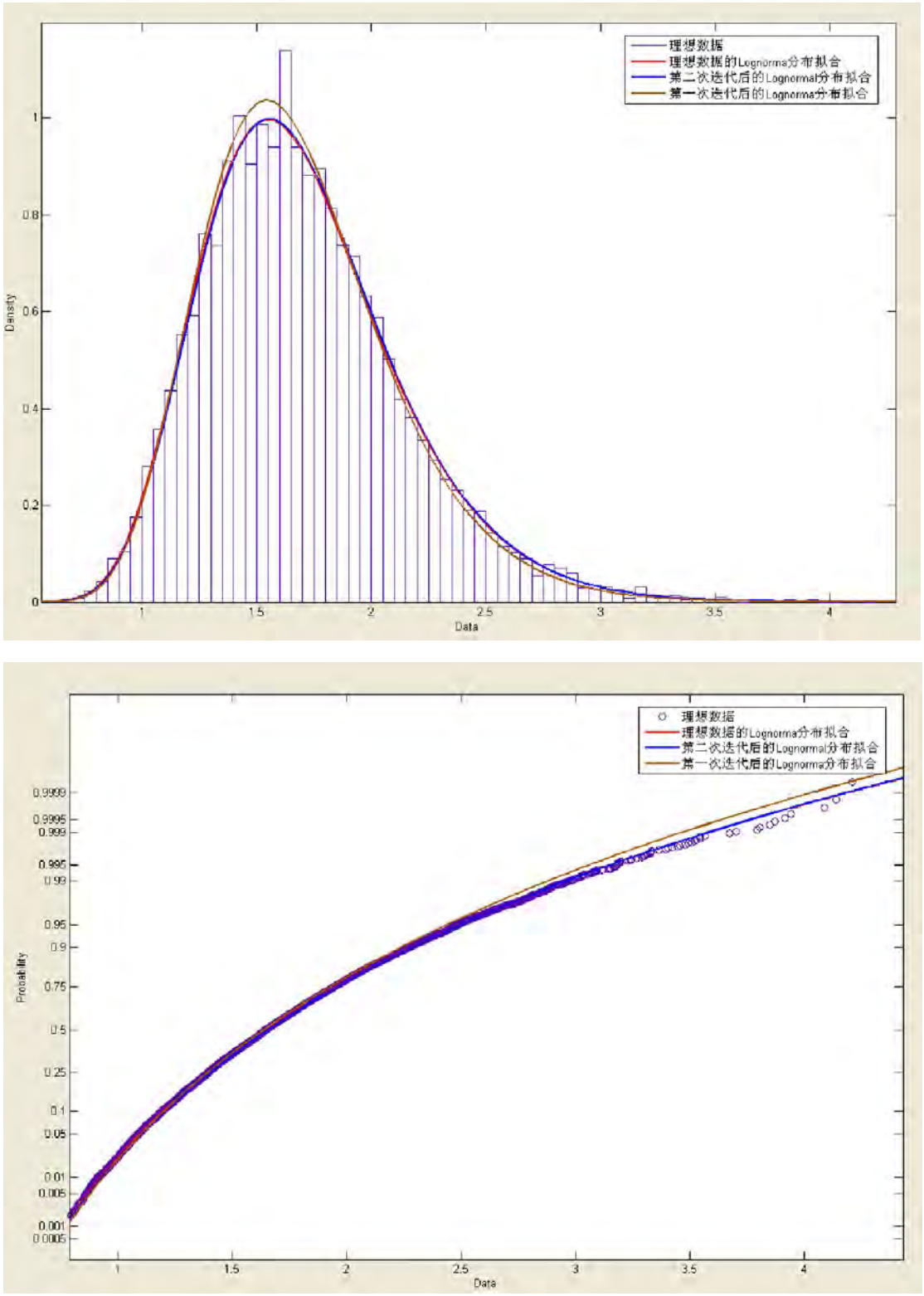


图 3.10 用 Lognormal 分布拟合时第二次迭代的结果

从上图可以看出，经过两次迭代，概率密度已经很好的逼近理想数据的 Lognormal 分布曲线。

仿真结果如下表所示：

表 3.3 迭代法仿真结果统计

名称	Log likelihood	Domain	Mean	Variance
理想数据	-5339.97	0 < y < Inf	1.70163	0.187185
第一次迭代	-4927.59	0 < y < Inf	1.68556	0.171308
第二次迭代	-5323.74	0 < y < Inf	1.70681	0.186405

我们的算法具有收敛速度快的优点，在第二次迭代后一般都能获得较好的概率密度函数的分布。

3.3 不同地区概率分布函数的组合问题

前面建立模型是考虑的样本情况是完全一致的，但实际情况却不是这么简单。比如说，全国的情况是由某些省、市的日常监测数据来估计的，这时也同样面临着两者的概率分布函数可能并不相同的问题。这里我们把调查数据看成是从若干个比较相近的总体的并集中有一定选择性地抽样所获得的数据，并用来估计这若干个有比较大共性的总体并集的概率分布函数，这时的问题相当于分层按比例抽样问题。此时总体的概率密度估计等于各层（提供监测数据的省、市）概率密度按其样本数所占比例的加权平均。即假设我们由 N 个省、市的检测数据来估计全国的情况，这 N 个检测数据的概率密度函数分别为 $f_1(x), f_2(x), \ldots f_N(x)$，该省、市的总人数分别为 $M_1, M_2, \ldots, M_N$，在这 N 个省、市的全部总人数为 $M = M_1 + M_2 + \ldots + M_N$。第 i 个省、市的权值定义为 $W_i = M_i/M$, $i = 1,2,\ldots,N$，则全国的概率密度函数估计为：

$$f(x) = \sum_{i=1}^{N} W_i f_i(x) = \sum_{i=1}^{N} \frac{M_i}{M} f_i(x) = \frac{1}{M} \sum_{i=1}^{N} M_i f_i(x) \qquad (3.7)$$

即我们这里将各个地区的抽样看成是分层抽样，由抽样调查的基本理论可知，对于分层抽样，用其各层估计均值的加权平均和来估计的总体的平均值的无偏估计。这一点我们已经在 2.3 节中进行了论述。

3.4 对样本频率分布曲线偏离预选分布很大的处理

以上讨论了利用有限的规则分布类型对样本的概率密度函数进行拟合，有限的规则分布类型难以满足实际应用的总体，以上的方法不能逾越给定先验分布类型、然后进行拟合优度检验这一基本思想，这本质上是给污染物分布强加了某种限制条件。

近年发展了基于密度演化理论的随机变量概率密度函数估计方法。在该方法中，不需要进行概率分布或概率密度函数类型的先验假定，可以直接根据基本数据给出概率密度函数的估计。

3.4.1 虚拟随机过程与密度演化方法

为了获取一组数据所来自的总体z的概率密度函数P (z)，可以构造一个虚拟随机过程

$$X(\tau) = \phi(Z, \tau) \qquad (3.8)$$

其中 τ 为虚拟时间参数。原则上，式（3.8）的函数形式只需使得z为 $X(\tau)$ 在 $\tau = \tau_c$

时刻的截口随机变量即可，亦即

$$Z = X(\tau)\big|_{\tau=\tau_c} = \phi(Z,\tau_c) \tag{3.9}$$

通常，为方便起见，式（3.8）的初始值可以取为

$$X(\tau)\big|_{\tau=0} = \phi(Z,\tau=0) = 0 \tag{3.10}$$

根据计算经验，式（3.8）取如下形式具有较好的效果：

$$\phi(Z,\tau) = Z\sin(\omega\tau), \tau_c = 1 \tag{3.11}$$

显然，当$\omega = 2.5\pi$时，式（3.11）满足式（3.9），式（3.10）中给出的条件。对式（3.8）关于τ求导可得$X(\tau)$的速率为：

$$\dot{X}(\tau) = \dot{\phi}(Z,\tau) \tag{3.12}$$

根据密度演化理论，(X，Z)构成的随机系统概率守恒，因而可以导出其联合概率密度函数$p_{XZ}(x,z,\tau)$满足的广义密度演化方程。

$$\frac{\partial p_{XZ}(x,z,\tau)}{\partial\tau} + \dot{\phi}(z,\tau)\frac{\partial p_{XZ}(x,z,\tau)}{\partial\tau} = 0 \tag{3.13}$$

由式（3.10）可知式（3.13）的初始条件为

$$p_{XZ}(x,z,\tau)\big|_{\tau=0} = \delta(x)p_Z(z) \tag{3.14}$$

边界条件可取$p_{XZ}(x,z,\tau)\big|_{x\to\infty} = 0$。

求解偏微分方程边值问题式（3.13）、式（3.14）可以给出联合概率密度函数$p_{XZ}(x,z,\tau)$，进而积分可得

$$p_X(x,\tau) = \int_{-\infty}^{\infty} p_{XZ}(x,z,\tau)dz \tag{3.15}$$

注意到式（3.9），即可获得随机变量z的概率密度函数

$$p_Z(z) = p_X(x=z,\tau)\big|_{\tau=\tau_c} \tag{3.16}$$

3.4.2 实施方法：

若有实测样本(Z_1,Z_2,Z_3,…,Z_{nsp})，欲以此为基本数据估计总体的概率密度函数。首先将方程式（3.13）关于z离散，并分别取已经获取的各个样本值，可得到一组偏微分方程

$$\frac{\partial p_{XZ}(x,z,\tau)}{\partial\tau} + \dot{\phi}(z,\tau)\frac{\partial p_{XZ}(x,z,\tau)}{\partial\tau} = 0 \quad (j=1,2,\cdots,n_{sp}) \tag{3.17}$$

在通常情况下，可以认为数据实测过程是独立重复进行的，因此，获取各个样本值的概率相同，故初始条件式（3.14）可取为

$$p_{XZ}(x,z,\tau)\big|_{\tau=0} = \delta(x)\frac{1}{n_{sp}|\Delta z|} \tag{3.18}$$

为方便计算，权值取$|\Delta z| = (z_{\max} - z_{\min})/n_{sp}$,(j=1,2,… ,$n_{sp}$)、在初始条件式（3.18）下求解方程式（3.17），可获得解答 $p_{XZ}(x,z,\tau)$(j=1,2,… , n_{sp})，进而，对式（3.15）进行数值积分可给出

$$p_X(x,\tau) = \sum_{j=1}^{n_{sp}} p_{XZ}(x,z_j,\tau)|\Delta z| \qquad (3.19)$$

结合式（3.18），式（3.19）不难看到，$|\Delta z|$的选取事实上对最终结果不产生任何影响，因此，在实际分析中，可直接取$|\Delta z|$=1。

式（3.17）可以采用差分格式求解。根据计算经验，在概率密度函数估计中，采用单边差分格式与具有TVD性质的修正的Lax-Wendrof格式构成的加权组合格式可取得较好的效果。具体实施步骤见图3.11。

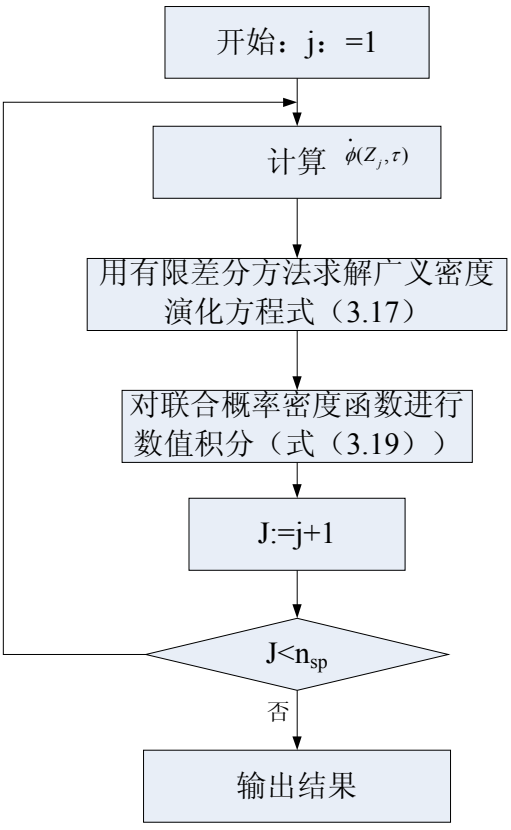


图3.11 概率密度函数估计的流程图

3.4.3 应用概率密度估计的效果图

图3.12概率密度函数与概率分布函数估计结果。为了比较，图3.12(a)还同时绘出了具有相同均值与标准差的极值I型分布的概率密度函数以及数据频率直方图。从图中不难看到，直接采用密度演化方法计算获得的概率密度函数曲线在趋势上与直方图最为符合。图3.12 (b)中同时绘出了经验分布函数、密度演化方法计算获得的概率分布函数和极值I型分布函数，从中可见，密度演化方法计算获得的概率分布函数与经验分布函数的偏差较小，在对可靠度问题影响较大的尾部

更是如此。

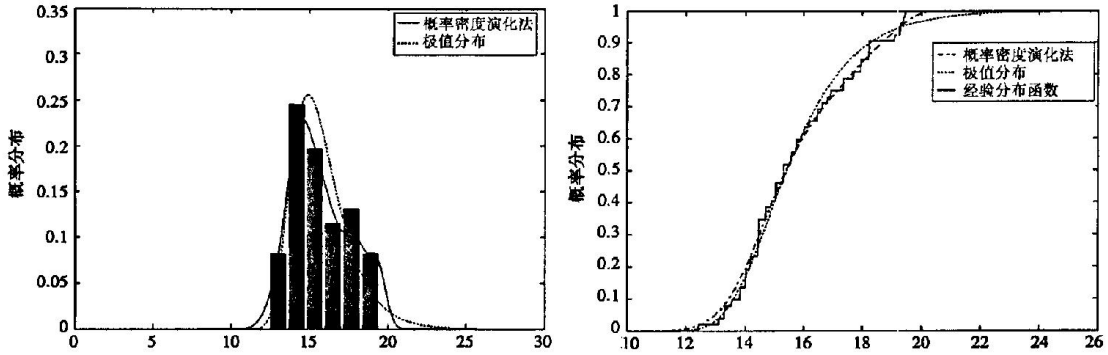


图 3.12 概率密度函数估计结果图

3.4.4 讨论

与传统的基于拟合优度检验的规则分布函数的统计方法比较，需要概率分布类型或密度函数类型先验分布的假定。研究表明：统计给出的结果是可信的。

虽然密度演化算法，具有不受先验分布的约束的优点，但是密度演化算法运算量大，而且受野值的影响很大，实际应用中要根据自己的需要来选择算法。

四、风险评估模型的建立

风险评估模型根据人群食物摄入量模型和污染物分布模型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（把每个人每天某种污染物摄入量看成是一个随机变量），从而能够对某一时刻食品安全风险做出评估。

4.1 风险评估模型的建立方法

风险评估模型就是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。模型的输入是抽样率很低的随机抽样数据，这两批数据通常是不配套的，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，这是建立风险评估模型需要首先解决的问题。在这里我们受原问题中给出的思路的启发，认为两批抽样调查是完全独立进行的。从实际问题考虑可以判断，这里给出的假设是合理的。在此假设下，风险模型的建立问题就转化为在相互独立的膳食模型和污染物分布模型给定的情况下来估计随机变量——每个人每天某种污染物摄入量的问题。假设对某种事物而言，某类人群对其一天的摄入量为随机变量 x，其分布为 $f_X(x)$（即为人群食品摄入量模型）；而每单位质量的该种食物中含有某类污染物的质量为 y，其服从分布 $f_Y(y)$（污染物分布模型）；则该人群对这种污染物的摄入量为随机变量 z，则有，$z=xy$，我们要求其分布 $f_Z(z)$（风险评估模型）。这一过程可以用图 4.1 中的左半部分来描述。

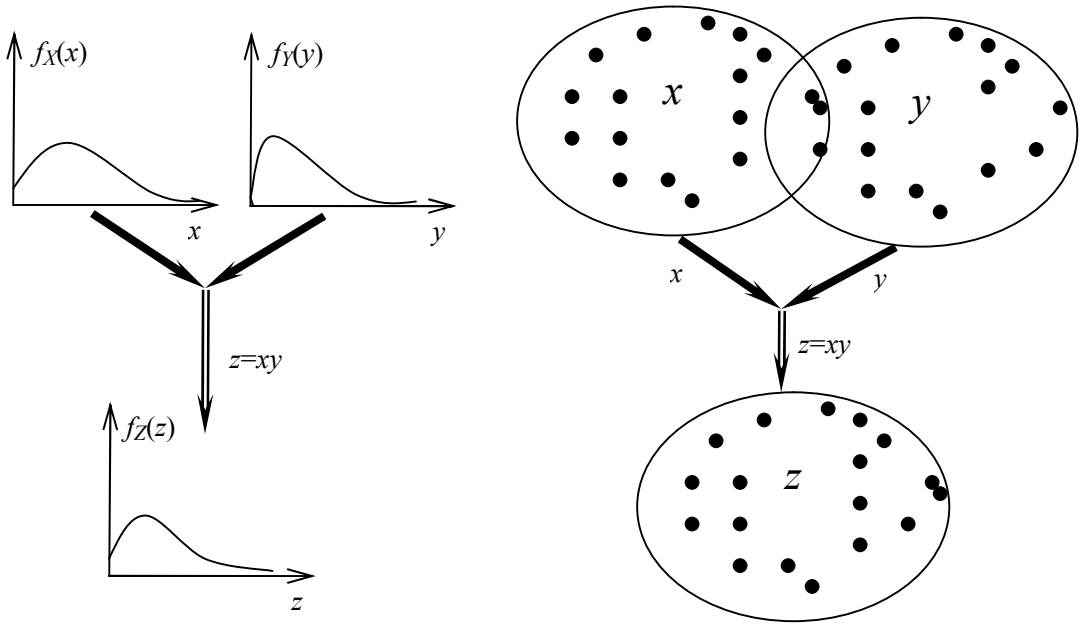


图 4.1 风险评估模型建立示意图

由 x 和 y 估计 z 实现过程如上图中的右半部分所示。由前面的两个环节，我们已经建立了食物摄入量（x）和污染物含量（y）的概率密度函数分布，而污染物的摄入量（z）为这两个量的乘积。假设 x，y，z 的累积分布函数（CDF）分别为 $F_X(x)$，$F_Y(y)$，$F_Z(z)$，则有：

$$F_Z(z)=\iint\limits_{xy\le z} f(x,y)dxdy=\iint\limits_{xy\le z} f_X(x)f_Y(y)dxdy \tag{4.1}$$

上式中的第二个等式是由我们假设两批抽样调查是完全独立进行而得到的。

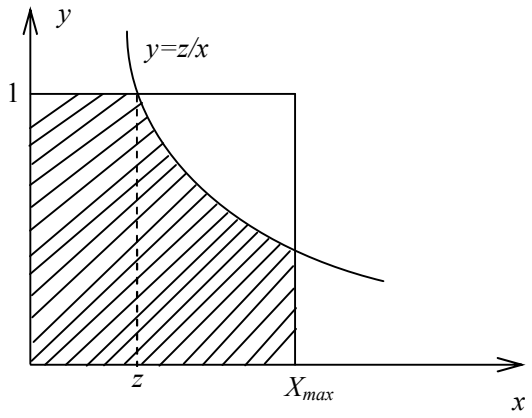


图 4.2 积分区间示意图

我们假设食品的摄入量的最大值为 X_{max}，而单位质量中污染物含量的最大值为单位质量 1，所以我们可得到积分区间，如图 4.2 所示，则积分可以展开为：

$$\begin{aligned}F_Z(z)&=\iint\limits_{xy\le z} f_X(x)f_Y(y)dxdy\\&=\int_0^1 f_Y(y)dy\int_0^z f_x(x)dx+\int_z^{X_{\max}} f_X(x)\int_0^{z/x} f_Y(y)dydx\end{aligned} \tag{4.2}$$

则概率密度函数为上式的导数：

$$f_Z(z) = F'_Z(z) = \frac{dF'_Z(z)}{dz} \qquad (4.3)$$

上式即为风险评估模型。

4.2 污染物摄入量 99.999%的右分位点精度的提高

风险评估模型对食品安全进行评估是通过人群食物摄入量模型和污染物分布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（把每个人每天某种污染物摄入量看成是一个随机变量），并将其与食品卫生安全部门的安全标准进行比较，如果右分位点小于安全标准，则认为食品卫生状况是安全的。其重点是对高暴露人群(即污染物摄入量比较大的人群)的监控上，而不仅是居民污染物的平均摄入量。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。所以右分位点估计的精度是非常重要的问题。

对概率密度函数的某一数值的右分位点的估计最直接的方法是：首先由样本数据估计函数的概率密度函数，然后计算使其积分上限使积分为要求的数值。这里如果用这种方法来估计污染物每天摄入量的 99.999%的右分位点，则有下面三个因素会影响到估计的精度：第一，目前还没有发现污染物摄入模型符合的分布，而只能用一些分布来近似，所以用样本来对其进行估计，不可避免的存在误差，当然用此概率密度函数来估计右分位点也会存在较大误差；第二，由建立污染物分布模型的过程可知，例行监测数据和符合性检验只在污染物超标时才会有数据结果，甚至其结果只是定性的，但我们都知道绝大多数食品还是安全的，超标的食品毕竟只占少数。偶然抽查数据和监测性检验数据可以提供样本数据，但其本身在整个检测中的比例非常小（2%），所以建立污染物分布模型时的样本数是很少的，而占食品绝大部分的安全食品的样本数会更少，由此也可以得出风险评估模型中样本数据量是比较少的，特别是安全食品部分的样本数会更少，而我们要估计右分位点恰恰需要用到安全食品部分的概率分布函数，所以会严重影响估计的精度；第三，在估计中用到了对概率密度函数的积分，在实现时只能用数值积分来近似，又会引人积分误差，而且其计算速度也会受影响。

针对直接估计 99.999%的右分位点方法的这些缺点，我们给出了一种新的估计 99.999%右分位点的方法。方法主要是在两个方面进行了改进来提高估计的精度：一是不估计随机变量的概率密度函数，而是用随机采样的样本来估计；二是我们不是依据 99.999%部分的样本数来估计右分位点，而是估计尾部的 0.001%的左分位点。

直接基于随机采样样本的右分位点的估计方法是基于 Mont Carlo 方法。这里首先介绍一下 Mont Carlo 方法的基本思想。

蒙特卡罗（Monte Carlo）原为地中海沿岸 Monaco 的一个城市，是世界闻名的大赌场。将 Monte Carlo 作为一种计算方法的命名固然已赋予了新的内容。然而，顾名思义，Monte Carlo 方法的随机采样特征在它的名字上得到了充分的反映。Monte Carlo 方法在数学上又被称为随机模拟（Random Simulation）方法、随机采样（Random Sampling）技术或统计试验（Statistical Testing）方法，它的基本思想是：根据要解决的具体问题建立适当的概率模型或随机过程，使其参数等于问题的解，然后通过对模型或过程的采样实验来计算所求参数的统计特征或

其近似值。如今，计算机的发展及其性能的提高使得 Monte Carlo 方法得到了广泛的应用。

Monte Carlo 方法提供了一种有效的用样本集来表示概率密度的途径，概率密度函数 $p(\boldsymbol{x})$ 基于样本集 $\{\boldsymbol{x}_i\}_{i=1,2,...N}$ 的表示形式为：

$$p(\boldsymbol{x}) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_i) \tag{4.4}$$

式中 $\delta(\bullet)$ 为 Kronecker delta 函数，$\{\boldsymbol{x}_i\}_{i=1,2,...N}$ 为来自于 $p(\boldsymbol{x})$ 的样本集，即 $\boldsymbol{x}_i \sim p(\boldsymbol{x})$。式（4.4）实为用采样点的疏密来表示概率密度。图 4.3 给出了一维概率密度函数 $p(x)$ 的 Monte Carlo 表示方法示意。

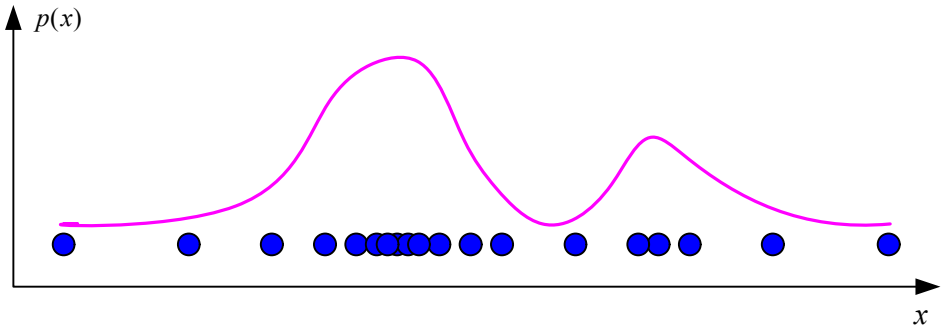


图 4.3 概率密度函数的 Monte Carlo 表示

然而，通常情况下，很难直接对 $p(\boldsymbol{x})$ 进行采样，重要性采样理论则摆脱了直接从真实概率分布进行采样的限制，它只需从一个易于实现采样的概率密度函数 $q(\boldsymbol{x})$ 进行采样，就理论而言，$q(\boldsymbol{x})$ 可以任意选择，通常把 $q(\boldsymbol{x})$ 称为重要性采样函数。利用重要性采样方法所得到的基于 Monte Carlo 样本集 $\{\boldsymbol{x}_i\}_{i=1,2,...N}$ 表示的概率密度函数 $p(\boldsymbol{x})$ 可以写成：

$$p(\boldsymbol{x}) \approx \sum_{i=1}^{N} w_i\delta(\boldsymbol{x}-\boldsymbol{x}_i) \tag{4.5}$$

$$w_i \propto \frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} \tag{4.6}$$

上式实为用一有权采样集 $\{\boldsymbol{x}_i, w_i\}_{i=1,2,...N}$ 来表示概率密度。图 4.4 给出了一个在一维情况下，用匀均分布作为重要性采样函数来表示概率密度函数 $p(x)$ 的示例，图中采样点的面积正比于其权重值。

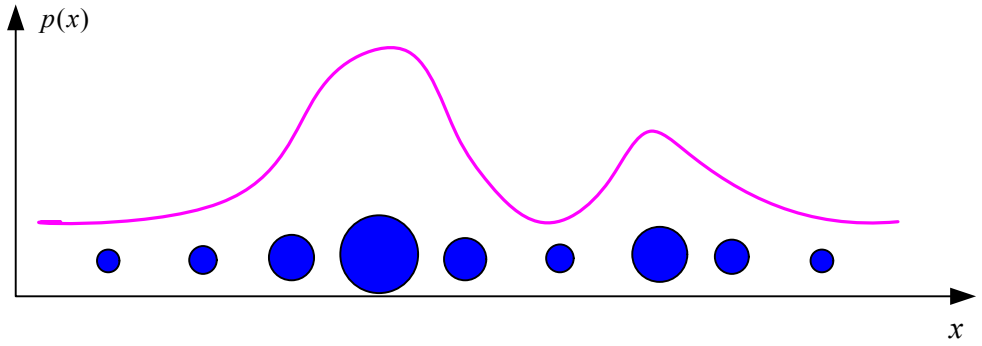


图 4.4 基于重要性采样的概率密度函数的 Monte Carlo 表示

在实际的工程应用中，我们常常需要对概率密度为 $p(\boldsymbol{x})$ 的某一随机变量 $\boldsymbol{x}=(x_1,x_2,...x_M)$ 或随机函数 $f(\boldsymbol{x})$ ）的数字特征进行估计。例如，在概率与统计理论中， $f(\boldsymbol{x})$ 的数学期望可以表示为：

$$E[f(\boldsymbol{x})]=\int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}=\int f(x_1,x_2,...x_N)p(x_1,x_2,...x_N)dx_1dx_2...dx_M \tag{4.7}$$

然而，式（4.7）中解析形式的数学期望在实际应用中一般难以直接计算。Monte Carlo 方法为解决这一问题提供了一种途径。基于 Monte Carlo 方法的数学期望可以表示成：

$$E[f(\boldsymbol{x})]\approx\frac{1}{N}\sum_{i=1}^{N}f(\boldsymbol{x}_i) \tag{4.8}$$

式中 $\{\boldsymbol{x}_i\}_{i=1,2,...N}\sim p(\boldsymbol{x})$，根据大数定理可得，当 $N\to\infty$ 时，$\frac{1}{N}\sum_{i=1}^{N}f(\boldsymbol{x}_i)$ 将趋于 $E[f(\boldsymbol{x})]$。

为了克服直接对 $p(\boldsymbol{x})$ 进行采样的困难，这里同样可以利用一重要性采样函数 $q(\boldsymbol{x})$ 进行采样得到样本集 $\{\boldsymbol{x}_i\}_{i=1,2,...N}\sim q(\boldsymbol{x})$，此时，基于 Monte Carlo 方法 $f(\boldsymbol{x})$ 的数学期望可以表示成：

$$E[f(\boldsymbol{x})]\approx\frac{1}{N}\sum_{i=1}^{N}f(\boldsymbol{x}_i)\frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}=\frac{1}{N}\sum_{i=1}^{N}f(\boldsymbol{x}_i)w_i \tag{4.9}$$

直接基于随机采样样本的右分位点的估计方法的基本原理如图 4.5 所示。

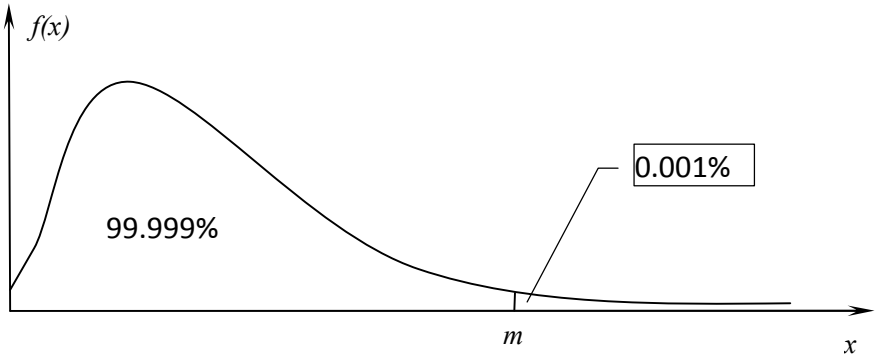


图 4.5 右分位点估计算法示意图

如图所示，图为随机变量 x 的概率密度函数曲线，我们要估计 m 值，使得 m 值左边的概率密度函数的积分占整个概率密度曲线积分面积的 99.999%。这里我们不是按照常规的方法先由抽样样本数据估计随机变量的概率密度函数曲线再积分。假设随机抽样得到的样本数为 M，先按照从小到大的顺序排列将抽样样本值；然后找出其中的左边的 99.999%M 个和其相邻右侧样本的分界线，并取其中值为 m 的估值，即假设随机变量样本在从小到大排列之后的从左边数第 99.999%M 个样本值为 m_1，m_2 是与 m_1 相邻且位于其右侧的样本，则 m 的估计值定义为：

$$\hat{m} = \frac{1}{2}\left(m_1 + m_2\right) \qquad (4.10)$$

为了保证估计的精度，对抽样的样本总数 M 有一定的要求。这里要求左边的面积为 99.999%，即为 0.99999，精度至少是 0.01，所以需要则需要的样本数至少为 10^7。

考虑到这里具体问题的情况，由前面建模的过程，我们知道虽然 m 左边的积分面积比其右边大得多，但左边的抽样样本数却是很有限的，因为这些数据很多都来自于复合性检验。而右边的部分虽然其所占面积很小（只有 0.001%），但是其中的样本数却非常的多，因为所有的检验都会给出 m 右边的准确测量值。所以我们很自然的可以想到，我们用 m 右边的数据来估计会得到更加准确的结果。所以这里我们的算法是：对于 M 个随机抽样样本值，我取右边 0.001%M 个样本，即从右边数的第 0.001%M 个样本，并找出与其相邻的左边的样本，去两者之间的均值作为 m 的估计。

五、小　结

论文主要建立了膳食暴露评估数学模型，主要是人群食物摄入量模型、污染物分布模型、风险评估模型三部分，并对建模过程中的一些理论和实际问题进行了探讨。针对膳食暴露评估数学模型的特点，提出了几种解决相应问题的算法，例如怎样根据随机变量取值大于某一值的部分统计数据估计出随机变量（或向量）的概率分布函数；两个不配套的抽样调查数据用什么方法去衔接使用才能达到理想的效果；调查数据中不相同的统计分类标准之间怎么转化；用不同地区的检测数据估计全国的情况时，两者概率密度函数分布不一致时的处理问题。我们从理论上在一定程度上找到了解决这些问题的方法，但并为找到实际的检测数据对所给算法进行检验，这将是我们今后需要研究的问题。

致　谢

由于我们并不了解食品安全卫生保障体系相关专业知识，因此我们在建模过程中参考了很多相关专业的资料和研究成果。对参考成果的引用在文中并未详细地的注明，而是将其列在论文的参考文献中。在此，我们对参考文献的作者表示衷心感谢！同时，我们衷心地感谢阅卷专家老师审阅我们的论文！

参考文献

[1]. Ralph L. Kodell, Seung-Ho Kang, James J. Chen, Statistical models of health risk due to microbial contamination of foods [J], Environmental and ecological statistical, 2002(9), 259-271.

[2]. L. Edler, K. Poirier, M. Dourson, J. Kleiner, Mathematical modeling and quantitative methods [J], Food and Chemical Toxicology, 2002(40), 283-326.

[3]. Arie H. Havelaar, Maarten J. Nauta, Jaap T. Jansen, Fine-tuning food safety objectives and risk assessment [J], International Journal of Food Microbiology, 2004(93), 11-29.

[4]. 俞纯权，宋廷山，山东省城镇劳动就业与社会保障抽样调查方案设计[J]，数理统计与管理，2005，24（2）；14-19.

[5]. 孙玉环，住房需求抽样调查方案设计及数据处理方法[J]，统计与决策（理论版），2007（6）；140-142.

[6]. 王亚文，肉及肉制品安全风险系数研究[D]：[硕士学位论文]，山东泰安：山东农业大学，2006.

[7]. 杨丽，朱运平，国际食品法典委员会（CAC）食品与饲料分类标准研究[J] 世界标准化与质量管理，2006（12）；36-37.

[8]. 杨义群，抽样调查与抽样检查——理论方法与应用[M], 北京：学苑出版社，1993.

[9]. 朱道元，吴诚鸥，秦伟良，多元统计分析与软件 SAS[M], 南京：东南大学出版社，1999.

[10]. 朱道元，数学建模精品案例[M], 南京东南大学出版社，1999.