全国第四届研究生数学建模竞赛

题号 A

题 目 基于自助法和核密度估计的膳食暴露评估模型

摘 要：

目前，我国食品安全形势严峻，迫切需要建立食品卫生安全保障体系。针对这个问题，本文建立了膳食暴露评估数学模型，并用对模型进行了检验和推广。该模型由以下四个模型组成：

模型Ⅰ：提出了一套完整的抽样调查评估模型。通过对不同影响因素的分析，给出了三种抽样方法和一些常用的调查手段，以及技术路线框图。

模型Ⅱ：按时令、地域、性别、年龄段和污染物含量标准区间分类，建立了某一时间段、某一地区特征人群每人每天食物摄入量的截尾正态分布模型。

模型Ⅲ：提出一种基于自助法的抽样处理方案。该方案在进行符合性检验，获得不合格食品的污染物含量的数据后，又进行偶然抽查的数据的基础上，实施再抽样以弥补小子样的不足。在此基础上，采用核密度估计方法建立污染物含量的分布模型，并且对如何合理的选取核函数进行了讨论。

模型Ⅳ：在模型Ⅱ和模型Ⅲ的基础上，建立了风险评估模型。考虑到数据不配套问题，提出一种基于模糊理论的解决方法。在此基础上，建立污染物摄取量模型，利用 Gauss-Legendre 求积公式，得到了 99.999%的右分位点。

全文对多个模型进行了仿真模拟。最后就预警标准进行讨论。

关键词：食品安全；膳食暴露；自助法；核密度估计

参赛密码 _____

（由组委会填写）

（一） 问题重述

1.1 基本情况

我国是一个拥有 13 亿人口的发展中国家，每天都在消费大量的各种食品，这批食品是由成千上万的食品加工厂、不可计数的小作坊、几亿农民生产出来的，并且经过较多的中间环节和长途运输后才为广大群众所消费，加之近年来我国经济发展迅速而环境治理没有能够完全跟上，以至环境污染形势十分严峻;而且随着我国进出口贸易的迅速增加，加上某些国外媒体的炒作，对外食品贸易中的矛盾也开始尖锐起来，因此建立包括食品卫生安全保障体系在内的公共安全应急机制是关系国计民生和对外贸易的重大而迫切的任务。我国建立食品卫生安全保障体系的时间还不长，根据国际上的热点和我国的国情，据初步估计，我国现阶段可能会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、有机磷、有机氯等实行监控，其他污染物的监控工作则待时机成熟后再推广。因此我国肯定也需要建立膳食暴露评估数学模型，建立我国自己的膳食模型，来在实施对污染物监控的同时，对公共食品卫生安全做出评估，并可以供领导决策时参考。

1.2 问题的由来

为了做出有效的膳食暴露评估数学模型根据现有资料看是分成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。其中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；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；风险评估模型则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（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作出评估。该模型的目标是保证绝大多数（99.999%以上）居民的食品安全,但重点却在对高暴露人群(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。如果用数学的语言严格地表述，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。如果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，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。当然这个右分位点相对于上述标准能够再向左一些，就能够保证更多居民的食品安全。

1.3 问题的要求

人群食物摄入量模型可以根据我国总膳食数据来建立，这批数据应该由调查人员入户调查获得，让调查人员事先进入被确定为调查对象的家庭，对居民家里的大米、面粉、食油、食盐，糖等全部食品进行称重并加以记录，几天后再来到这户居民家中并将他们家里的大米、面粉、食油、食盐，糖等全部食品称重，将两次结果相减就可以得出这户居民在这几天中所消费的各类食品的总量，并对没有称重的食品，如蔬菜、水等的消费情况也进行登记；再将调查所得的全部统计数据汇总就得到我国总膳食数据的抽样结果。由于这项调查工作量太大，如果实行普查，其工作量甚至超过全国人口普查，故而只可能在全国几亿户家庭中随机抽取几千户，至多几万户进行一次性调查。因此如何设计抽样调查方案使调查结

果能尽量反映全国的实际情况，调查结果的数据使用起来效果比较理想，同时使调查的全部工作量在可以承受的范围内，是一项困难的任务。这项工作的另一个难点在于中国居民消费的食品种类比其他国家居民消费的食品种类复杂得多，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，在实际调查过程中进行如此详细地分类，其调查工作量太大，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。这项工作的第三个任务是要用通过万分之一（甚至更小）的抽样率得到的数据建立起全国比较准确的人群食品摄入量模型，因此要确定合理的技术路线，充分利用从其他一切渠道可以获得的信息，可以并且应该建立不止一个这样的模型以满足各方面的需求。

污染物分布模型主要是根据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，前者的结果可能只是定性的，而后者检测的结果精度高）和市场上各类食品的流通量，此外还包括进出口口岸的检测数据来估计市场上各种食品的污染物含量。建立这个模型同样有以下几个难点：第一个难点是这里的数据也是抽样率很低的随机抽样的数据，否则工作量太大，且无法满足监测时间方面的要求，问题是应该怎样充分利用这批数据去建立模型？第二个难点是由于食品的季节性、区域性、多样性特点，日常监测无法获得详细的、完整的分类数据，问题是如何利用这些数据尽量提高模型的精度？第三个难点是由于监测时间方面的要求和经费的限制，在日常检测时往往采用比较快捷的检测方法，即符合性检验，其缺点是当检测项目的检测结果是安全时就不再精确测量污染物具体的含量了，而笼统地用“未检出”作为检测结果。这对判断这批食品是否安全而言是完全满足要求的，但作为污染物分布模型的输入而言，如果“未检出”全部当成零来计算就一定会产生比较大的误差，因此一定要改进。用数学的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（如通过大约占数据总量 2%的偶然抽查数据所获得的小于等于同一数值的部分样本数据）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，对可能出现的食品安全事件给出预警。首先这个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模型？第二个难点是两个模型的数据分类也很可能不配套，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（历史数据无法按现在的要求进行修改），在模型中如何妥善处理这样的问题？第三个难点是这个模型要求给出全体居民某项污染物摄入量的 99.999%的右分位点，如何提高它的精度？

因此我们要建立食品卫生安全保障体系除按美国和欧盟的方法需要建立三个数学模型外，还希望提出有创造性的技术路线（如食品卫生安全保障体系数学模型的全新的整体设计方案，调查数据的总体结构和调查方案，如何综合利用一切有用信息等等），同时迫切需要研究、解决大量的理论问题。这里的难点是对比较多的情况都掌握得并不十分清楚，因此不妨做出一些必要、合理的假设，并在此基础上进行详细的分析。

（二）基本假设

1. 假设抽样地区的家庭组成一致，均有老人、成人和孩子，且男、女比例相当；
2. 假设在同一季节，同一地区内，同一性别，同一年龄段的人群，饮食习惯相似；
3. 假设在全国范围内，食物的供求相当；
4. 假设在整个分析过程中，没有任何自然灾害及政治变动等突发事件影响；
5. 假设每个季节，每个地区，各个不同特性的人群每人每天对各类食物的摄入量服从截尾正态分布。

（三）符号设定

n——样本容量

$Z_{\alpha/2}$——置信水平为$1-\alpha$的标准正态分布的上分位点

p——总体估计的差异性

E——p与p的估计值的绝对误差不超过的值

X——某季节，某地区，特征人群每人每天摄取的食物量

x——某季节，某地区，特征人群每人每天摄取食物量的观测值

Y——某季节，某地区的污染物量

y——某季节，某地区的污染物量的观测值

Z——某季节，某地区，特征人群每人每天摄取的食物含污染物含量

z——某季节，某地区，特征人群每人每天摄取食物中污染物含量的观测值

（四）问题分析

为了建立膳食暴露评估数学模型，首先需要进行合理抽样，为了简化模型，在建立技术路线模型时，只对某时间，某地区的饮食情况进行调查。据有关资料介绍可以获得五种抽样方案，在对这几种方案的可实现性，可操作性分析后，我们提出了一种可操作的抽样调查方案，在此基础上，建立人群食物摄入量模型，污染物分布模型和风险评估模型。

考虑到不同季节，不同地区，不同特征的人群的饮食习惯差异，食物摄入量模型应分别讨论。但由于没有真实的数据，我们假定每个季节，每个地区，各个不同特征人群每人每天对各类食物的摄入量服从不同参数的截尾正态分布。

对于污染物分布模型，由于抽样率极低，我们采取自助法处理数据，这样可以尽量利用当前少量数据实现对总体概率分布的估计。由于对总体的信息知道的太少，我们利用核密度估计方法估计总体的分布密度，由于本问题的随机变量都是非负的，可以选用截断密度函数作为核函数。

为了在实施对污染物监控的同时，对公共食品卫生安全做出评估，并可以供领导决策时参考。在已建立的三个模型基础上，还需建立风险评估模型。此时把食物中的污染物含量和食物的摄入量看成是独立的，建立两者的联合分布模型，来合理、有效的评估食物中的污染物含量。本模型的一个难点就是特征对象或数据的不配套问题，受模糊近似理论的启示，可以合理地把不配套问题转化成配套问题来解决。

（五）模型的建立与问题解决

5.1 模型Ⅰ：抽样调查技术路线模型

这里规划的抽样调查技术路线模型大致分为：明确调查目的，熟悉调查总体，

分析调查条件，确定样本容量，选择抽样方法，分析评估样本情况这几个环节。其中确定样本容量是非常关键的一步。

5.1.1 样本容量

样本容量是抽样的老问题，其大小取决于对监测指标的估计精度要求以及所用抽样方案的设计效应。若总体容量较小，则可以采用普查，但在实际中我们所调查的总体总是容量很大的。显然不能用普查，而大量事实证明，当总体样本很大时，采用抽样所做出的估计比较符合实际情况，但是考虑工作量的负担，需要一个最小的子样容量，使之满足有效原则，可测量原则和简单原则。现在对样本容量取多大比较认同的有两种方法：

其一，固定比例法，就是调查样本容量是总体的固定比例；

其二，置信区间法，就是运用差异性区间、样本分布以及平均数标准或百分率标准误差等，通过公式计算算出样本容量大小，

计算公式为：

$$n = \frac{Z_{\alpha/2}^2 \times p \times (1-p)}{E^2} \tag{1}$$

5.1.2 抽样方案的设计

在确定该子样容量的基础上，根据统计学抽样原理，在参考了其它大型社会调查的抽样方案后，给出一个合理的抽样模型，即技术设计路线模型。基于问题的需要，提出如下三个重要的随机抽样方法：

（1）自助法重复抽样：

在母体中，抽取样本容量为 n 的随机样本后，将其放回母体，同样再次抽取，直到达到要求。其优点是可以减少随机抽取中的误差，使其更接近母体的分布特征。

由于本次调查要求最小的样本容量，对某一季节，某一地区的特征人群进行重复抽样，可以在一定工作量限制下提高精度，减小误差。

（2）分段抽样：

先在一母体中，抽取 n 个单位随机样本，再由单位随机样本中抽出 m 个子单位，就子单位进行调查，称为二段抽样。若续从子单位中抽取更小单位进行调查，称为三段抽样。三段以上，则称为多段调查。其优点是可以提高精度，剔除一些坏值。

在本次调查中，先按照调查对象人数多少成比例的随机等距抽样法，再针对某种特性进行进一步随机抽样。也可以根据不同情况，反复抽样。

（3）分层抽样：

先设立目的及某种分类标准分为若干组或若干类，此组类称为层，然后将母群体之各个体分别编入相当层中，再由各层中抽样选取适量样本的方法。其基础有赖抽样设计者的经验和判断。理想上分层的数目越多越好。这是由于层数越多，每层的样本单位越相似，样本估计值的精确度越高。但从成本考虑，层数不宜过高。

由于本次调查所考虑分类因素较多，为研究某一因素的影响，可以基于该因素将调查对象适当分类，再分别对每一类进行抽样。如为研究年龄的影响，可将调查的人群按不同年龄段分为老人、中年人、青少年、儿童和婴儿。

5.1.3 常规食物摄取量调查方法

常规食物摄取量调查主要是针对上述的三种随机抽样方法提出的，具体情况如下。

（1）称重法：

称重法就是在某一个伙食单位（家庭）一段时间内，运用日常的各种测量工具，对各类食物食用进行称重，计算每伙食单位（家庭）中平均每人每日的各类食物摄入量。

步骤：

①准确记录每餐各类食物名称；

②准确称量：摄入前食物的质量和食物的剩余量；

③计算食物的总摄入量＝摄入前食物的质量－食物的剩余量；

④综合求得平均每人每日的食物消耗量；

⑤将调查期间所消耗的食物按品种分类、并计算平均每人每日的各类食物摄入量。

此方法的特点是：通常由调查对象或看护者在一定时间内完成，对调查人员的技术要求高，而且被调查对象必须很好的合作配合。其缺点在于在外就餐的消耗食物汇报的准确性差。食物记录过程可能影响或改变其日常的饮食模式；随记录天数的增加，记录的准确性可能降低；而且经常发生高报，低报等现象，大量的高报，低报估计多发生在一些特定人群（如肥胖人群）；长期记录时会给被调查者带来较多的麻烦，不适合大规模、长时期调查。

针对这些问题，可用称重法得出的数值作为基准，再与其它统计方法得出的数值作比较，相对数值误差为±25%均可作为有效数据，置信水平定值在95%。

（2）查账法：

查账法是由被调查对象或研究者记录一定时期内的食物消耗总量，研究者通过调查这些记录并根据同一时期进餐人次数，计算平均每人每天各类食物的平均摄入量。

步骤：

①准确记录每餐各类食物名称和每天进餐人次数；

②准确称量：摄入前食物的质量和食物的剩余量；

③计算所有人次食物总摄入量=摄入前食物质量—食物剩余量；

④综合求得平均每人每日的食物消耗量；

⑤将调查期间所消耗的食物按品种分类，并计算平均每人每日的各类食物摄入量。

此方法适用于有详细伙食账目的集体单位,也可用于家庭。在帐目精确和每餐用膳人数统计确实的情况下相当准确，并可调查较长时期的膳食状况，适用于全年四个季度的调查，调查的手续较简便，所费的人力少，且易于为膳食管理人员掌握，使调查单位能定期的自行调查计算，作为改进膳食质量的参考。

（3）24 小时回顾法：

24 小时回顾法是通过对受试者尽可能准确地回顾调查前一段时间，如前一目至数日的食物消耗量。询问调查前一天的食物消耗情况，获得平均每人每天各类食物的平均摄入量。

步骤：

①询问每餐各类食物名称和每天食物摄入量；

②准确记录询问的总人数，食物名称及每天食物摄入量；

③综合求得平均每人每日的食物消耗量

④将调查期间所消耗的食物按品种分类、并计算平均每人每日的各类食物摄入量。

此方法的特点是：在实际中，虽然很省时，省力，简单易行。但偏差比较大，所以建议调查数据采用多量、多次，来减小误差。

（4）化学分析法：

化学分析法就是通过实验室化学分析方法，测定调查对象在一定时间内所摄入食品的能量和营养素的数量及质量。

其主要目的往往不仅是收集食物消耗量，而且要在实验室中测定调查对象一日内全部食物的营养成分，准确地获得各种营养素的摄入量。最准确的收集样品方法是双份饭菜法：制作两份完全相同的饭菜，一份供调查对象食用，另一份作为分析样品。其特点是：收集的数据比较精确、可靠，但是需要耗费大量的人力、物力。

（5）食物频率法：

食物频率法是将食物分为若干类，分别询问调查对象过去一段时间或一年里，每类食物的摄入频率及平均每次食用量。是估计被调查者在指定的一段时期内吃某些食物的频率的方法。

其主要采用随机问卷调查的方法来进行，问卷包括内容根据实际的需要而定，一般应包括食物名单，食物频数，问卷调查法可调查过去某一非严格指定时间范围内(如一周,一月,一年等)各种食物消耗的频率及数量,进而获得进食量绝对值,根据此方法得到的食物和营养素摄取量将个体划分为不同等级,食物频数问卷调查法既能获得个体在过去一段时间对象的膳食模式及饮食习惯,又能量化食物及营养素摄入量。其特点是：灵活、简单、数据量大，但相对而言数据精度不是很好，该方法一般用于研究食物和营养素摄入量与疾病的关系(表 1 综合给出了五种方案的比较)。

表 1 五种方案的比较

	优点	缺点	应用
称重法	准确	费时、费力 不适用大规模	家庭、个人、团体
查账法	简单易行， 省时、人、物	时间短不够准确， 代表性有影响	账目清楚的机关 部队学校
24 小时回顾法	简单易行， 省时、人、物	主观，不太 准确，回忆偏倚	家庭、个人
化学分析法	更准确	费时、力、财	科研、治疗膳食
食物频率法	应答者负担轻，应答率高，经济、方便； 可调查长期	量化不准确（偏高）， 遗漏	个人，膳食习惯与某些慢性疾病的关系

5.1.4 食品摄取量调查技术路线模型

通过比较分析得知，化学分析法成本较高，不推荐使用，食物频率法主要用来长期调查一些慢性疾病。故这里基于自助法重复抽样，分段抽样和分层抽样，综合称重法，查账法和 24 小时回顾法，提出如下食品摄取量抽样调查方案。

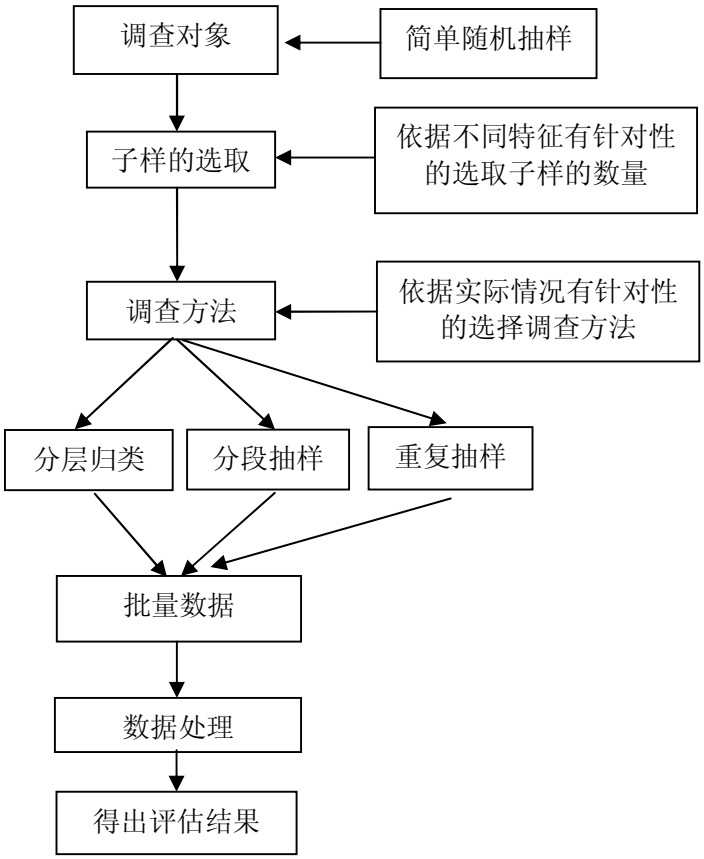


图 1 食品摄入量抽样调查方案

简要说明：

（1）简单随机抽样调查对象，该方法具有健全的理论基础，是一种客观而实用的方法，具有普遍性，在市场调查中经常使用。

（2）为使抽样数据反映现实情况，我们依据不同特征（不同地区、不同性别、不同年龄、不同季节、不同劳动强度、不同经济收入等人群各类食品的一天摄入量）有针对性的选取子样的容量。

（3）基于合理的子样容量，依据具体的实际情况（操作性、技术及能力等因素），选取有效、合理的调查方法。对于农村、城市工作不稳定和长期流动人口，可以采用 24 小时回顾法和食物频率法进行调查，对于城市工作稳定人员、政府机关、部队和学校可以考虑查帐法，但对于经济幅度变化快和某些偏远地区可以采用称重法、化学分析法，调查对象少也可以说明问题。

（4）为保证抽样结果的真实、可靠性，且能估计出调查对象的特性，采用先分层归类，再兼顾分段抽样、重复抽样。可得到批量数据。

（5）对这些批量数据进行整理，分析，并做出评估。

5.2 模型Ⅱ：人群食物摄入量模型

5.2.1 分类标准的设定

(1)季节集合：{春，夏，秋，冬}；

(2)区域集合：{华中，华北，华南，华东，华西，西北部，西南部，东南部，东北部}；

(3)性别集合：{男，女}；

(4)年龄段集合：{婴儿，儿童，青少年，中年，老人}；

(5)按各污染物将食品划分成以下集合：{铅；标准较低类，标准中等类，标准较高类}，{镉；标准较低类，标准中等类，标准较高类}，{有机氯；标准较低类，标准中等类，标准较高类}，{有机磷；标准较低类，标准中等类，标准较高类}；

这里所谓“标准”指的是国家标准。对食品这样分类主要是由于中国居民消费的食品种类比其他国家居民消费的种类要复杂得多，细分将达数千种以上，在调查过程中如果详细分类，调查工作量大，而如果随意粗糙分类，工作量虽然小，但影响精度，而且意义不大。因此我们根据各种污染物在这些食品中的分布情况进行区间分类。比如要得出铅(Pb)在各种食物中的分布，我们可以按国家标准把食品大致分为三分类(表 3)(单位：mg/kg)。

表 2 根据铅含量对食品按区间模糊分类

分类区间	(0，0.2]	(0.2，0.5]	(0.5，2]
类别	a	b	c

由于每一类包含百种食品，可以简单地将各类别表示出来，即 $a=${标准较低类，如蔬菜,水果,蛋类}，$b=$ {标准中等类，如肉类，鱼类，奶制品，小食品}，$c=$ {标准较高类，如饮料，罐头，皮蛋}[9]。

经过这样分类，可以实施具体情况抽样，例如有这样一个集合{铅，标准较高类；春，华东，男，老人} 是指对春季华东地区男性高龄人群摄取含铅标准较高类食物的抽样调查。

5.2.2 截尾正态分布函数假设

由于缺失详细调查数据，我们无法对所要求的人群充分调查，所以根据经验我们将其分布近似正态分布，但由于正态随机变量可以取负值，而这与根据实际情况不符，为此我们在正态分布的基础上，构造一个函数来近似人群食物摄入量的密度函数，即截尾正态分布函数 $f(x)$.

定义 1 若随机变量 X 有密度函数

$$f(x)=\begin{cases}0, & \text{其 他}\\ \dfrac{1}{\sqrt{2\pi}\,c\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, & x\geq 0\end{cases} \tag{2}$$

则称 X 服从截尾正态分布，其中 $c,\sigma>0,-\infty<\mu<\infty$ 。

5.2.3 人群食物摄入量模型

现在来验证所构造的函数是否能较好的接近实际的人群食物摄入量的密度分布函数。

对于人群食物摄入量密度分布函数 $g(x)$ 来说（ x 是随机变量，本模型中指食物摄入量），应满足以下几个条件：

$$\begin{cases}\mathrm{g}(x)=0, x<0\\ \lim\limits_{x\to 0}\mathrm{g}(x)\to 0, \lim\limits_{x\to\infty}\mathrm{g}(x)\to 0\\ \mathrm{g}(x)>0, x>0\\ \int_{-\infty}^{\infty}\mathrm{g}(x)dx=1\end{cases} \tag{3}$$

可以验证，上面所定义的截尾正态分布函数 $f(x)$ 在一定的条件下可以满足

以上条件，来确定 c 值；

由 $\int_0^{\infty} f(x)dx = 1$ 可得：

$$\int_0^{\infty} \frac{1}{\sqrt{2\pi}c\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\frac{\mu}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}c} e^{-\frac{t^2}{2}} dt = \frac{1}{c}(1-\Phi(-\frac{\mu}{\sigma})) = 1 \tag{4}$$

其中：$t = \frac{x-\mu}{\sigma}$

$$\therefore c = 1-\Phi(-\frac{\mu}{\sigma}) = \Phi(\frac{\mu}{\sigma}) \tag{5}$$

其中，$\Phi(x)$ 是服从标准正态分布的，即 $\Phi(x) \sim N(0,1)$，而且要求 σ 与 $e^{\frac{\mu^2}{2\sigma^2}}$ 的乘积足够的大。

综上所述，我们就可以确定出人群食物摄入量分布密度函数，表达式为：

$$f(t) = \begin{cases} 0, & \text{其他} \\ \dfrac{1}{\sqrt{2\pi}\Phi(\frac{\mu}{\sigma})\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}, & t \geq 0 \end{cases} \tag{6}$$

从表达式可以看出密度函数仅由 μ, σ 确定的。

图 2 给出某季节，某地区，特征人群对于某类食物每人每天的摄入量分布函数，其中：$\mu = 2.5$，$\sigma = 1.3$。

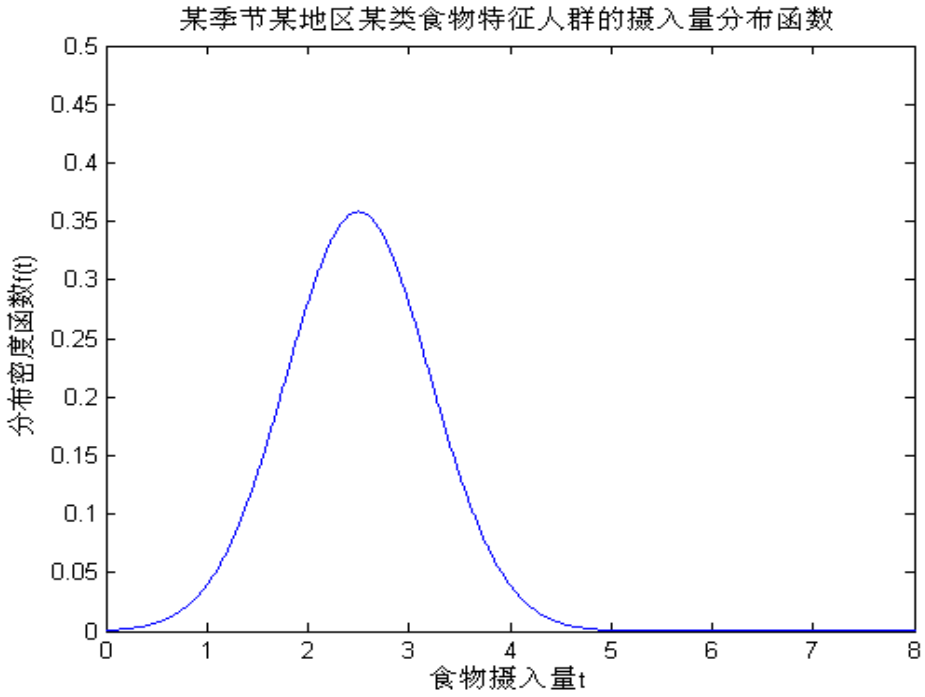


图 2 特征人群对于某类食物每人每天的摄入量分布函数

5.3 模型III：污染物分布模型

污染物分布模型的建立首先依赖于市场上食品的检测数据。但考虑到工作量、经费和监测时间等方面限制，往往只能获得抽样率很低的随机数据，如何充分合理利用现有数据是一个难点。根据题目给出的信息，我们可以重点考虑符合

性检验和偶然抽查这两方面的数据，由符合性检验得到不合格食品的污染物含量的数据，而合格食品的数据可通过在合格样本中再抽取少量样本来获得。污染物分布模型建立的另一个难点是没有具体数据，而且我们对其可能的分布密度函数的信息了解太少，在这种情况下通常采取非参数估计的方法来求得分布函数，如核密度估计就是一种优良的非参数密度估计方法。

5.3.1 随机抽样数据的获得与处理方案

自然界中食品的污染物含量总是以一定规律分布的，即它的概率密度函数是存在的，而且是未知的，只能通过大量的随机抽样数据去估计，因此抽样数据的获得及其处理对模型的建立非常重要。

随机抽样数据的来源主要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。综合考虑工作量、成本和监测时间等方面因素，我们仅采取符合性检验和偶然抽查数据来获得估计分布函数所需的样本数据。其过程如下：首先，对样本容量为 n 的某类食品中的某种污染物含量进行符合性检验，按照国家规定的卫生标准，将样本分为合格与不合格两类，其样本数量分别为 n_1、n_2，并得到不合格样本的污染物含量数据。对于合格的样本，由于没定量信息，我们必须对其重新抽样获得数据。一般而言，合格样本总是占多数，为减小工作量和缩减成本，我们在合格样本中，只偶然抽查占样本总量2%的样本数。并且，为不改变原样本中合格与不合格样本数之间的比例($n_1 : n_2$)，对这 $n \cdot 2\%$ 个合格样本，我们采用一种统计推断方法，即自助法(Bootstrapping)[5]。自助法是一种再抽样(resampling) 统计方法，其要点是：①假定观察值便是总体；②由这一假定的总体抽取样本, 即再抽样。由原始数据经过再抽样所获得的与原始数据集含量相等的样本称为再抽样样本(resamples) 或自助样本。如果将由原始数据集计算所得的统计量称为观察统计量的话，那么由再抽样样本计算所得的统计量称为自助统计量。自助法的关键是自助统计量与观察统计量间的关系, 就如同观察统计量与真值间的关系，可表示为：

自助统计量 ∶∶ 观察统计量 ⇔ 观察统计量 ∶∶ 真值

其中，“∶∶”表示二者间的抽样关系, “⇔”表示等价于。也就是说, 通过对自助统计量的研究,就可以了解有关观察统计量与真值的偏离情况。其中的再抽样是有返还的抽样方式。设问题中2% 的偶然数据为 m 个观察值，那么自助样本可按如下步骤获得：

①将每一观察值写在纸签上；

②将所有纸签放在一个盒子中；

③混匀。抽取一个纸签, 记下其上的观察值；

④放回盒子中, 混匀, 重新抽取；

⑤重复步骤③和④ m 次, 便可得到一个自助样本。

重复上述抽样过程 B 次, 便可得到 Bm 个自助样本。假设“未检出”的食品数量占整个检测数据总量 $\eta\%$，即

$$\eta\% = \frac{\text{“未检出”食品数量}}{\text{检测数据总量}}$$

则可以得出这个抽样过程的次数

$$B = \left\lceil \frac{\eta\%}{2\%} \right\rceil \tag{7}$$

通过自助法我们可以最大程度地利用极少数的抽样样本(观察统计量)得到自助样本，最终使样本的合格率恢复到原来的值。在此基础上，我们可以通过一

些非参数估计的方法对概率密度函数进行估计。

5.3.2 模型的建立

由给定样本点集合求解随机变量的分布密度函数问题是概率统计学的基本问题之一，解决这一问题的方法包括参数估计和非参数估计。在参数判别分析中，人们需要假定作为判别依据的、随机取值的数据样本在各个可能的类别中都服从特定的分布、经验和理论说明，参数模型的这种基本假定与实际的物理模型之间常常存在较大的差距，这些方法并非总能取得令人满意的结果。由于上述缺陷，Rosenblatt 和 Parzen 提出了一种非参数估计方法，即核密度估计方法[6]。由于核密度估计方法不利用有关数据分布的先验知识，对数据分布不附加任何假定，是一种从数据样本本身出发研究数据分布特征的方法，因而，在统计学理论和应用领域均受到高度的重视。

根据已有经验，我们仅知道污染物分布不是正态分布而且很可能是一种偏态分布，但是它是否符合我们已经掌握的一些偏态分布，如 χ^2 分布、F 分布、对数正态分布等等，尚无定论。由于我国是一个人口众多，地理情况极为复杂的发展中国家，我们不能采用发达国家的经验模型来衡量我国的污染物含量分布情况，为此我们采用核密度函数去估计污染物含量分布。核估计的定义如下：

定义 2 设 $Y_1, Y_2, \cdots, Y_n$ 是从一维总体 Y 中抽出的独立同分布的样本，其服从的分布密度函数为 $f(y)$，$y \in \mathbb{R}$。定义函数

$$\hat{f}_n(y) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_n}\right) \tag{8}$$

为总体未知密度函数 $f(y)$ 的一个核估计。其中 $K(\cdot)$ 为核函数，h_n 称为窗宽或光滑函数，且满足 $\lim_{n\to\infty} h_n = 0$。

由定义可知，分布密度函数的核密度估计不仅与给定的样本点集合有关，还与核函数的选择和带宽参数的选择有关．其中，带宽参数 h_n 控制了在求点 h_n 处的近似密度时不同距离样本点对点密度的影响程度，$K(\bullet)$ 须为 $(-\infty, \infty)$ 上的Borel可测函数。

常用的核函数有：高斯核函数(Gaussian Kernel)、Epanechnikov 核函数和 Biweight 核函数。图 3 所示的用高斯核来估计密度函数的一个例子，从图中可以看出核估计的物理意义是：在获得样本数据的前提下，通过有限个已知的标准分布函数（核）的叠加，来近似未知分布密度函数。

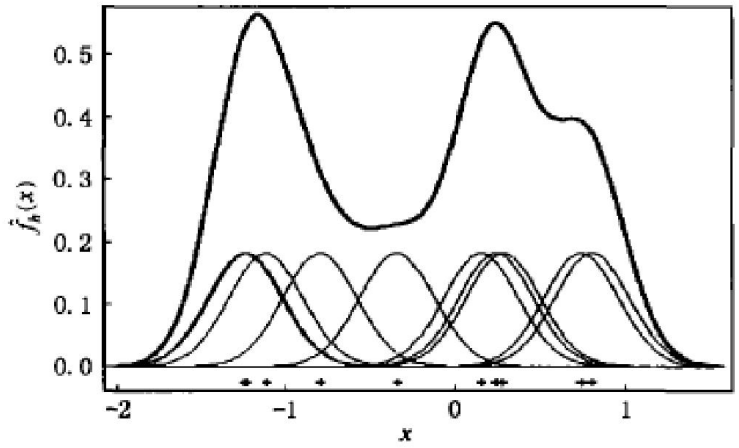


图 3　由 10 个点描绘的高斯核估计曲线

针对本题中的具体问题，我们先只考虑在某一地区、某一个时间段内，某一类食品中含某一种污染物的量，至于其他地区，不同时间段，各类食品含各种污染物的量的情况，完全可以通过相关的样本数据由该模型同理求出其分布。

设某一地区食品分为 N_1 类，污染物分为 N_2 类，定义随机变量 Y 为：某类食品中的某种污染物的含量。设 $f(y)$ 为 Y 的概率密度函数，利用核密度函数，则可将其估计为：

$$\hat{f}(y)=\frac{1}{nh_n}\sum_{i=1}^{n}K(\frac{y-Y_i}{h_n}) \tag{9}$$

其中 Y_i 表示随机变量 Y 的样本数据中的第 k 个数值。这里我们将污染物的含量视为相对量，即食品单位质量中所含污染物的质量，其单位为 mg / kg 。

建立核估计模型的首要任务是选取核函数。理论上核 K 可以灵活选取，但从实用上看，将 K 选为密度函数比较合理。这是因为待估计的是分布密度，最好估计量 f 本身也是密度函数。并且当 K 为密度时，容易验证 f 满足概率密度函数的归一化条件。因此我们将核函数的选取范围缩小到已知的密度分布函数之中。但是，目前常用的作为核函数（如前面所举的几个例子）的分布函数大多都是呈坐标原点对称的函数，而在本题中，污染物的含量显然是非负值，并且污染物的分布应该是一种偏态分布，如果用上述函数作为核函数，一方面不满足随机变量的定义域条件；另一方面用对称核函数叠加，不利于得到偏态分布函数。所以，经过综合考虑，我们选取如下具有偏态性质的对数正态分布函数

$$K(t)=\frac{1}{\sqrt{2\pi}\sigma t}\exp[-\frac{1}{2\sigma^2}(\ln t-\mu)^2]\,,\ t>0 \tag{10}$$

作为核估计的核函数。因此，最终我们建立的污染物分布模型如下：

$$\hat{f}(y)=\frac{1}{nh_n}\sum_{i=1}^{n}\frac{h_n}{\sqrt{2\pi}\sigma(y-Y_i)}\exp[-\frac{1}{2\sigma^2}(\ln\frac{y-Y_i}{h_n}-\mu)^2]\,,\ y>0 \tag{11}$$

5.3.3 精度分析

5.3.3.1 大样本性质

由于中国居民消费的食品种类比其他国家居民消费的种类要复杂得多，细分将达数千种以上，所以即使我们按 5.2.1 将某污染物按区间模糊分类，仍无法缓解由于种类多工作量大的问题，对食品含量的抽查还是需要成千上万次。我们知道当样本很大趋于无穷，即 $n\to\infty$ 时，$\hat{f}_n$ 一定条件下也会趋于原始概率密度 f 。

定理 1（L_r 收敛）[8] 设 f 为有界概率密度，$\hat{f}_n$ 在 $(-\infty,\infty)$ 在上有界且处处连续，$K(u)$ 为概率密度，满足条件：

(1) $K(u)$ 在 $(-\infty,\infty)$ 上有界；

(2) $a_1=\int_{-\infty}^{\infty}uK(u)du<\infty$，$a_2=\int_{-\infty}^{\infty}u^2K(u)du<\infty$，$\cdots$；

(3) $\lim\limits_{|u|\to\infty}\left|uK(u)\right|\le M$

则当 $h>0$，$h\to 0$，$nh\to\infty\,(n\to\infty)$ 时有 $\int_{-\infty}^{\infty}\left|\hat{f}_n-f\right|^r dx\to 0\,(n\to\infty),a.s.\,(r\ge 1)$。

同样这也适用于定义域 $(0,\infty)$ 的情形，只要我们抽样调查的样本容量足够大，比如大于十万或者更多，核密度估计就能最大程度地贴近原始概率密度。

5.3.3.2 窗宽调节

我们用 $T_n(y)\triangleq T_n(y;Y_1,\ldots,Y_n)$ 表示基于样本 $Y_1,\ldots,Y_n$，对未知密度函数 f 的任

一估计。由于$T_n(y)$既同样本有关，又是考察点的函数，Rosenblatt(1956 年)提出一种广泛使用的整体测度，即积分均方误差(MISE)：

$$\begin{aligned} MISE(T_n) &= \int E[T_n(y) - f(y)]^2 dy \\ &= \int [ET_n(y) - f(y)]^2 dy + \int Var(T_n(y)) dy \end{aligned} \tag{12}$$

因而

$$MISE = \text{积分偏差平方和＋积分方差.}$$

对于核估计来说，应该选择h_n使得相应的核估计的*MISE*达到最小，但在实际问题中，如何选择最佳窗宽是个很难处理的问题。

为便于计算及理论分析，下面我们分别导出估计偏差及方差的渐进表达式。为简单计，当$n \to \infty$，$h_n \to 0$，由控制收敛定理可得*MISE*的渐进公式：

$$MISE \approx \frac{1}{4} h_n^4 a_2^2 \int [f''(y)]^2 dy + (nh_n)^{-1} \int K^2(u) du \tag{13}$$

其中$K(u)$满足：

$$a_1 = \int uK(u) du = 0 \text{；} \quad a_2 = \int_{-\infty}^{\infty} u^2 K(u) du \neq 0 \tag{14}$$

f具有二阶有界连续导数，再对h求极小，得到渐进的最佳窗宽：

$$h_{opt} = a_2^{-2/5} [\int K^2(u) du]^{1/5} [\int (f''(y)^2 dy)]^{-1/5} n^{-1/5} \tag{15}$$

这表明最佳渐进窗宽随n增大以$n^{-1/5}$的速度趋于零，但最佳窗宽仍然很难求得。文献[7]提及一些方法，如参数参照法，极大似然交叉证实法，最小二乘交叉证实检验法等等，但无论使用哪种方法，得到的窗宽都已经失去“最佳窗宽”的含义了。而如果h_n选的过大$\hat{f}_n$对f有较大的平滑，使得f的某些特性被掩盖，h_n选得过小$\hat{f}_n$又波动较大。经验选择$h_n = n^{-\alpha}$，$\alpha > 0$。只要α准确，即使提供给我们的抽样点少，我们也可以叫准确地估计出污染物分布模型。换句话说，适当地选择 窗宽h_n可以弥补样本量少的不足。

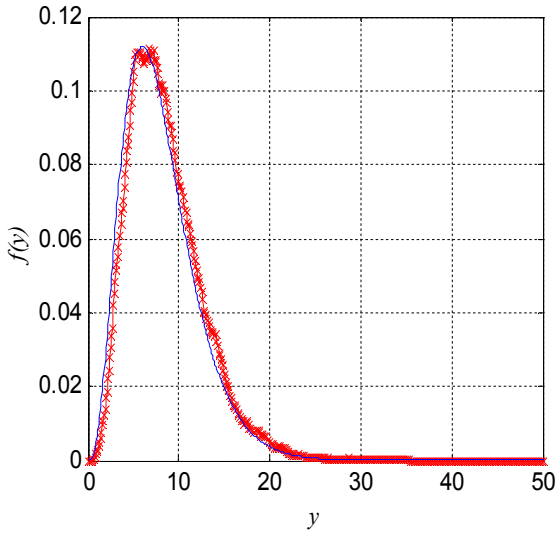
5.3.3.3 μ和σ的选取

这两个参数是由核函数引入的，它们决定了核函数的形状。而根据已有研究表明，核函数的形状对分布估计的精度是有影响的。因此，应该根据样本数据合理选择μ和σ的值，同样可以提高估计精度，弥补样本需求大的不足。

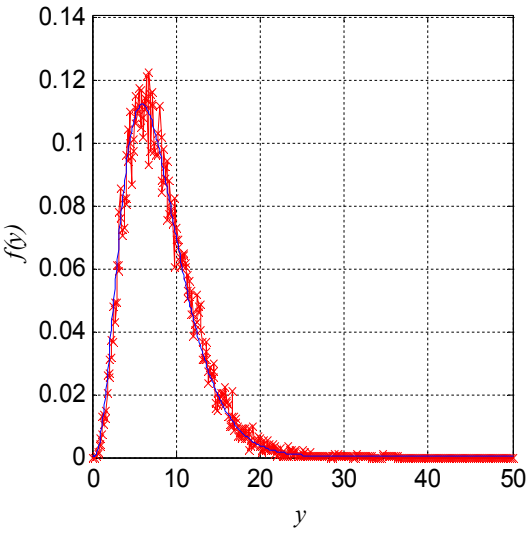
5.3.4 模型的检验

在不知道污染物分布函数和没有充足有效数据的前提下，验证我们所建模型的正确性是相当困难的。这里我们提出一种思路：先假设污染物分布是某种已知的密度函数$f'(y)$（如χ^2分布、F 分布等），然后生成一定量的服从该密度函数的随机数据，再将这此数据代入到所建模型中，求出$f(y)$的具体表达式，最后比较$f(y)$和$f'(y)$，检验两者在误差允许范围内，是否能较好吻合。

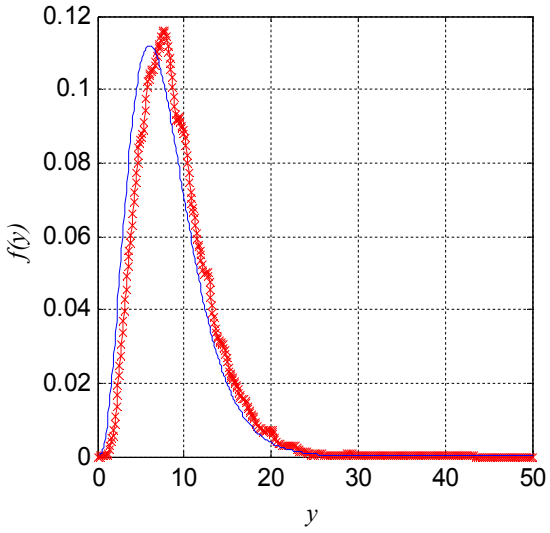
由于缺少数据，我们不妨假设污染物含量服从χ^2分布，利用计算机生成10000 个服从χ^2分布的数据，然后利用数据估计出分布函数，其结果如下图所示。



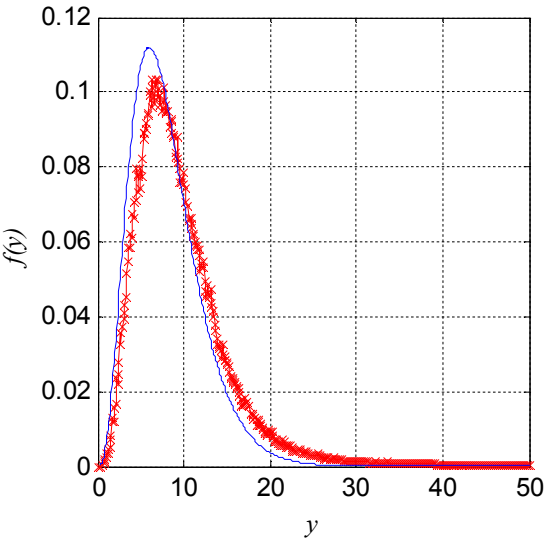
(a) $\alpha=0.01$，$\mu=-1$，$\sigma=0.3$ 时的结果



（b）$\alpha=0.08$，$\mu=-1$，$\sigma=0.3$ 时的结果



(c) $\alpha=0.01$，$\mu=0$，$\sigma=0.3$ 时的结果



(d) $\alpha=0.01$，$\mu=-1$，$\sigma=2$ 时的结果

图 4 对数正态分布核密度估计的检验结果

由上图可以看到，以对数正态分布为核函数的核密度估计整体上是可行的，各参数虽然对分布有一定影响，但如果在有足够大的样本的条件下，这种影响不明显，故针对本题只要调查充分，我们还是可以非常精确地描述出污染物的具体分布情况的。另外核函数的选取也对整个分布模型的估计影响较大，因为高斯核的定义域太过宽泛，这显然不符合本题题意，这里也说明采用对数正态分布核对于本模型的分布估计是有效的。而且选取的窗宽是否合理也会对估计分布的光滑度有一定的影响，h_n 越小越粗糙。

5.4 模型Ⅵ：风险评估模型

风险评估模型就是利用前两个模型的结果对全国、某地区、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。但有一个问题是这两批数据很有可能是不匹配的，即食品摄入量模型和污染物分布模型所针对的不是一类特征人群，并且两者所针对的食品很可能不是一致的。

5.4.1 Gauss-Legendre 求积公式

为合理、有效的评估食物中的污染物含量，通过分析，建立食物的摄入量与污染物分布的联合分布模型。为简单起见，只考虑两种模型数据匹配的情况。该模型的建立主要依据 Gauss 型求积公式，在给出该公式之前，我们先定义正交多项式和 Legendre 多项式。

定义 3：对于任意实变量 x，满足：

$$\psi_{n+1}(x)=A\omega_{n+1}(x)=A\prod_{i=0}^{n}(x-x_i), \tag{16}$$

其中，A 为常数，x_i 为 $\psi_{n+1}(x)$ 的零点，则称 $\psi_{n+1}(x)$ 为 $n+1$ 正交多项式。

定义 4：对于任意实变量 x，满足：

$$P_n(x)=\frac{1}{2^n n!}\frac{d^n}{dx^n}\left[(x^2-1)^n\right] \tag{17}$$

则称 $P_n(x)$ 为 n 次 Legendre 多项式。

至此，我们给出 Gauss-Legendre 求积公式：

设[a,b]=[-1,1]；$\omega(x)\equiv 1$，则正交多项式族 $\{\psi_k\}$ 为 Legendre 多项式族，则下式：

$$\int_a^b \omega(x)f(x)dx=\sum_{i=0}^{n}A_i f(x_i)+\frac{f^{(2n+2)}(n)}{(2n+2)!}\int_a^b \omega(x)\omega_{n+1}^2(x)dx \tag{18}$$

可化为：

$$\int_{-1}^{1} f(x)dx=\sum_{i=0}^{n}A_i f(x_i)+\frac{f^{(2n+2)}(n)}{(2n+2)!}\int_{-1}^{1} \omega_{n+1}^2(x)dx \tag{19}$$

其中 $x_1, x_2, \ldots x_n$ 为 n+1 次 Legendre 多项式的零点，且

$$A_i=\int_{-1}^{1} l_{n,j}^2(x)dx \tag{20}$$

它们的数值如下表：

表 3　Gauss-Legendre 求积节点和系数

n	x_i	A_i
0	0	2
1	± 0.5773502692	1
2	± 0.7745966692 0	5/9 8/9
3	± 0.8611363116 ± 0.3399810436	0.3478548451 0.6521451549
4	± 0.9061798459 ± 0.5384693101 0	0.2369268851 0.4786286705 0.568888889

上式称为 Gauss-Legendre 求积公式，简称 Gauss 公式。

对于[a,b]区间上 $\omega(x)\equiv 1$ 的情形，可作变量

$$x=\frac{a+b}{2}+\frac{b-a}{2}t \tag{21}$$

将变量化为 $t\in[-1,1]$，再利用上式计算。

5.4.2 风险评估模型的构建

根据模型Ⅱ的讨论，可以得到食物摄取量服从截尾正态分布，即

$$f_1(x)=\frac{1}{c\sqrt{2\pi}\sigma_1}e^{(-\frac{(x-\mu_1)^2}{2\sigma_1^2})}, \tag{22}$$

而再根据模型Ⅲ的讨论，得出污染物含量密度分布函数——高斯核密度函数，即

$$f_2(y)=\frac{1}{nh_n}\sum_{i=1}^{n}\frac{h_n}{\sqrt{2\pi}\sigma(y-Y_i)}e^{[-\frac{1}{2\sigma^2}(\ln\frac{y-Y_i}{h_n}-\mu)^2]} \tag{23}$$

再根据风险评估提出的要求，根据在食物中的污染物含量和食物的摄入量与污染物的分布的关系，故提出建立两者的联合分布模型。经分析知，二者相互独立。于是，对于每人每天摄取污染物含量 z，其随机变量 $Z=XY$，其概率分布如下：

$$P\{Z\le z\}=P\{XY\le z\}=\iint\limits_{xy\le z}f(x,y)dxdy=\int_0^{x_{\max}}f_1(x)dx\int_0^{\frac{z}{x}}f_2(y)dy \tag{24}$$

作代换：令 $t=(2y-\frac{z}{x})/\frac{z}{x}$

$$\begin{aligned}P\{Z\le z\}&=\int_0^{x_{\max}}f_1(x)\frac{z}{2x}\int_{-1}^{1}f_2(\frac{\frac{z}{x}t+\frac{z}{x}}{2})dtdx\\&=\frac{z}{2}\int_0^{x_{\max}}\frac{f_1(x)}{x}\left[f_2(\frac{z}{2x}(t_0+1))+f_2(\frac{z}{2x}(t_1+1))\right]dx\end{aligned} \tag{25}$$

再令 $s=(2x-x_{\max})/x_{\max}$

$$=\frac{z}{2}\int_{-1}^{1}\frac{x_{\max}}{2}\frac{f_1(\frac{x_{\max}s+x_{\max}}{2})}{\frac{x_{\max}s+x_{\max}}{2}}[f_2(\frac{z(t_0+1)}{2\frac{x_{\max}s+x_{\max}}{2}})+f_2(\frac{z(t_1+1)}{2\frac{x_{\max}s+x_{\max}}{2}})]ds \tag{26}$$

这里，由 Gauss-Legendre 求积公式知：当 $n=1$ 时，$A_0=A_1=1$，$t_0=-0.5773502692$，$t_1=0.5773502692$

$$\begin{aligned}\text{原式}&=\frac{z}{x_{\max}}\frac{x_{\max}}{2}\int_{-1}^{1}\frac{f_1(\frac{x_{\max}}{2}(s+1))}{s+1}[f_2(\frac{z}{x_{\max}}\frac{t_0+1}{s+1})+f_2(\frac{z}{x_{\max}}\frac{t_1+1}{s+1})]ds\\&=\frac{z}{2}\left\{\frac{f_1(\frac{x_{\max}}{2}(s_0+1))}{s_0+1}\left[f_2(\frac{z}{x_{\max}}\frac{t_0+1}{s_0+1})+f_2(\frac{z}{x_{\max}}\frac{t_1+1}{s_0+1})\right]+\frac{f_2(\frac{x_{\max}}{2}(s_1+1))}{s_1+1}\right.\\&\left.\left[f_2(\frac{z}{x_{\max}}\frac{t_0+1}{s_1+1})+f_2(\frac{z}{x_{\max}}\frac{t_1+1}{s_1+1})\right]\right\}\end{aligned} \tag{27}$$

$\because t_0=s_0, t_1=s_1$

$$\therefore \text{原式}=\frac{z}{2}\left\{\frac{f_1(\frac{x_{\max}}{2}(s_0+1))}{s_0+1}\left[f_2(\frac{z}{x_{\max}})+f_2(\frac{z}{x_{\max}}\frac{t_1+1}{s_0+1})\right]+\frac{f_1(\frac{x_{\max}}{2}(s_1+1))}{s_1+1}\right.$$

$$\left.\left[f_2(\frac{z}{x_{\max}}\frac{t_0+1}{s_1+1})+f_2(\frac{z}{x_{\max}})\right]\right\} \tag{28}$$

令：$c_1=\frac{1}{x_{\max}}$， (29)

$$c_2=\frac{t_1+1}{x_{\max}(s_0+1)}=\frac{t_1+1}{x_{\max}(t_0+1)}=\frac{t_1+1}{(t_0+1)}c_1 \tag{30}$$

$$c_3=\frac{1}{x_{\max}}\frac{t_0+1}{s_1+1}=\frac{1}{x_{\max}}\frac{t_0+1}{t_1+1}=\frac{t_0+1}{t_1+1}C_1, \tag{31}$$

$$c_5=\frac{x_{\max}}{2}(s_0+1), \tag{32}$$

$$c_6=\frac{x_{\max}}{2}(s_1+1). \tag{33}$$

则：$\text{原式}=\frac{z}{2}\{\frac{f_1(c_5)}{s_0+1}[f_2(c_1z)+f_2(c_2z)]+\frac{f_1(c_6)}{s_1+1}[f_2(c_3z)+f_2(c_1z)]\}$ (34)

$\because$ $f_1(x)=\frac{1}{\sqrt{2\pi}c\sigma_x}e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}}$

$$f_2(y)=\frac{1}{nh_n}\sum_{i=1}^{n}\frac{h_n}{\sqrt{2\pi}\sigma(y-Y_i)}e^{[-\frac{1}{2\sigma^2}(\ln\frac{y-Y_i}{h_n}-\mu)^2]} \tag{35}$$

$$\text{原式}=\frac{z}{2}\left\{\frac{f_1(c_5)}{s_0+1}\left[\frac{1}{nh_n}\sum_{i=1}^{n}\frac{h_n}{\sqrt{2\pi}\sigma(y-Y_i)}e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_1z-Y_i}{h_n}-\mu)^2]}+\sum_{i=1}^{n}\frac{h_n}{\sqrt{2\pi}\sigma(y-Y_i)}e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_2z-Y_i}{h_n}-\mu)^2]}\right]+\right.$$

$$\left.\frac{f_1(c_6)}{s_1+1}\left[\frac{1}{nh_n}\sum_{i=1}^{n}\frac{h_n}{\sqrt{2\pi}\sigma(y-Y_i)}e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_3z-Y_i}{h_n}-\mu)^2]}+\sum_{i=1}^{n}\frac{h_n}{\sqrt{2\pi}\sigma(y-Y_i)}e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_1z-Y_i}{h_n}-\mu)^2]}\right]\right. \tag{36}$$

令：$2\sqrt{2\pi}nh_n=l$ (37)

$$=\frac{z}{l}\left\{\frac{f_1(c_5)}{s_0+1}\left[\sum_{i=1}^{n}\frac{h_n}{\sqrt{2\pi}\sigma(y-Y_i)}e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_2z-Y_i}{h_n}-\mu)^2]}+\sum_{i=1}^{n}\frac{h_n}{\sqrt{2\pi}\sigma(y-Y_i)}e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_1z-Y_i}{h_n}-\mu)^2]}\right]\right.$$

$$\left.+\frac{f_1(c_6)}{s_1+1}\left[\sum_{i=1}^{n}\frac{h_n}{\sqrt{2\pi}\sigma(y-Y_i)}e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_3z-Y_i}{h_n}-\mu)^2]}+\sum_{i=1}^{n}\frac{h_n}{\sqrt{2\pi}\sigma(y-Y_i)}e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_1z-Y_i}{h_n}-\mu)^2]}\right]\right\} \tag{38}$$

令：$b1=\frac{f_1(c_5)}{(s_0+1)l}, b2=\frac{f_1(c_6)}{(s_1+1)l}$ (39)

$$= b_1 z \sum_{i=1}^{n} \frac{h_n}{\sqrt{2\pi}\sigma(y - Y_i)} e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_2 z - Y_i}{h_n} - \mu)^2]} + b_1 z \sum_{i=1}^{n} \frac{h_n}{\sqrt{2\pi}\sigma(y - Y_i)} e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_1 z - Y_i}{h_n} - \mu)^2]}$$

$$+ b_2 z \sum_{i=1}^{n} \frac{h_n}{\sqrt{2\pi}\sigma(y - Y_i)} e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_3 z - Y_i}{h_n} - \mu)^2]} + b_2 z \sum_{i=1}^{n} \frac{h_n}{\sqrt{2\pi}\sigma(y - Y_i)} e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_1 z - Y_i}{h_n} - \mu)^2]} \tag{40}$$

$$= (b_1 + b_2) z \sum_{i=1}^{n} \frac{h_n}{\sqrt{2\pi}\sigma(y - Y_i)} e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_1 z - Y_i}{h_n} - \mu)^2]} + b_1 z \sum_{i=1}^{n} \frac{h_n}{\sqrt{2\pi}\sigma(y - Y_i)} e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_2 z - Y_i}{h_n} - \mu)^2]}$$

$$+ b_2 z \sum_{i=1}^{n} \frac{h_n}{\sqrt{2\pi}\sigma(y - Y_i)} e^{[-\frac{1}{2\sigma^2}(\ln\frac{c_3 z - Y_i}{h_n} - \mu)^2]} \tag{41}$$

为使计算结果更高，可对区间$[0, \frac{z}{x}]$作k等分，区间$[0, x_{\max}]$作l等分，得：

$$= \frac{z}{2k} \sum_{m=0}^{l-1} \sum_{j=0}^{k-1} \left\{ \frac{f_1(\frac{s_0 + 2m + 1}{2} \Delta_l)}{s_0 + 2m + 1} \left[f_2(\frac{(t_0 + 2j + 1)z}{(t_0 + 2m + 1)k\Delta_l}) + f_2(\frac{(t_0 + 2j + 1)z}{(t_0 + 2m + 1)k\Delta_l}) \right] + \right.$$

$$\left. \frac{f_1(\frac{s_1 + 2m + 1}{2} \Delta_l)}{s_1 + 2m + 1} \left[f_2(\frac{(t_0 + 2j + 1)z}{(s_1 + 2m + 1)k\Delta_l}) + f_2(\frac{(t_1 + 2j + 1)z}{(s_1 + 2m + 1)k\Delta_l}) \right] \right\} \tag{42}$$

$$= 1 - \alpha \tag{43}$$

由题意知，$\alpha = 0.00001$.解上述方程，得即为其上分位点。

利用 MATLAB 软件生成某种左偏态随机数，再利用二分法思想编程，，即可找到z_α。

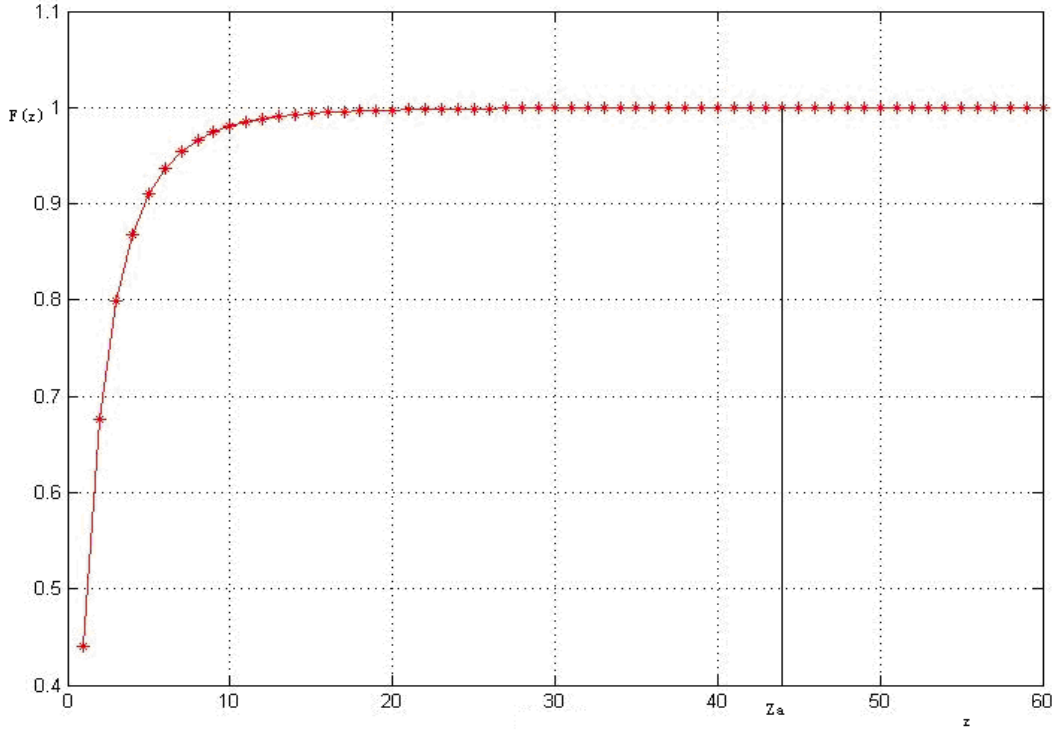


图 5　z 的分布函数 $F(z)$

5.4.3 数据、特征人群不匹配分析

5.4.1 和 5.4.2 的有关推导都是针对匹配模型而言，当两种模型的人群、数据不匹配时，我们可以采用模糊近似的方法来描述匹配度。用数学模型描述如下：当 X，Y 的实际数据不匹配时：假设我们已知 X，但不知道对应匹配的 Y，知道其它的 n 个 $Y_1,\ldots,Y_n$，则根据模糊数学的相关理论，首先我们给出一个 Y 与 $Y_i(i=1,\cdots,n)$ 的相似评判标准（7 等分标准）

表 4 相似评判标准

	极相似	很相似	较相似	相似	较不相似	很不相似	极不相似
分值	10	9	7	5	3	1	0

再求出权重向量 $A=(a_{ij})_{1\times n}, i=1,2,.....n$， a_i 的确定可用下表确定：

表 5

C	10	9	7	5	3	1	0	b_i	a_i
Y_1	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	$\sum_{j=1}^{7} c_j a_{1j}$	$b_1 / \sum_{j=1}^{7} a_{1j}$
Y_2	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	a_{27}	$\sum_{j=1}^{7} c_j a_{2j}$	$b_2 / \sum_{j=1}^{7} a_{2j}$
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
Y_n	a_{n1}	a_{n2}	a_{n3}	a_{n4}	a_{n5}	a_{n6}	a_{n7}	$\sum_{j=1}^{7} c_j a_{nj}$	$b_n / \sum_{j=1}^{7} a_{nj}$

则我们可以计算出 $\hat{Y} = A(Y_1 Y_2 ... Y_n) = \dfrac{\sum_{i=1}^{n} a_i Y_i}{\sum_{i=1}^{n} a_i}$

可用来近似的表示与 X 相匹配的 Y，即 $Y=\hat{Y}$，进而运用上面已经给出的匹配模型进行解决。该方法实际上就是将用 $Y_1,\ldots,Y_n$ 线性线性组合成与 X 近似匹配的 $\hat{Y}$，可以应用至两组模型人群的相似性度量和数据的相似性度量，通过线性组合后，直接用于 5.4.1 和 5.4.2，得到分位点。

通过以上分析，我们可以得出某食物中含有的某污染物含量右分位点，与国家预警标准比较，确定针对某种食物含某种污染物是否预警。

5.4.3 模型评价及推广

1．模型的优点：

(1)本文所建立的模型均有成熟的理论基础，又有相应的专业软件的仿真支持，可信度高；

(2)模型III中采用了自助法抽样，在保证精度的条件下，大大减小了工作量和成本, 而且采用核估计的方法来建立污染物分布模型，更具一般性；
(3)在模型Ⅳ中，为解决数据不配套问题，本文应用了模糊数学理论，这为解决这类问题提供了一种较好的思路；
(4)推广性强，本文建立的模型不仅可对污染物的评估，而且可以推广到解决其他类似问题。
2．模型的缺点：
(1)模型检验的数据通过计算机生成，与实际情况可能有一定出入；
(2)由于人群食物摄入量模型是我们根据现有资料作出的假设，对风险评估模型的精度会产生一些影响。
3．模型的推广

由于模型III是一般性估计，只要根据实际情况选择合适的核函数，完全可以推广到对其他未知分布的密度函数估计中。在实际中，时间也是影响污染物分布密度的一个重要因素，所以，如果考虑时间因素，则污染物含量分布就是一个一个随机过程来处理，那么将会得出更优的数学模型。

对于模型Ⅳ来说，处理匹配问题时，我们先从简单入手，解决匹配时的估计，进而通过一定的转化，将不匹配的转化成匹配的求解。这种思路可以很好的运用到实际中去，对于比较复杂的问题而言，若是直接解答没有思路，则可以转化思考的方向，先从最简单的方向入手，把其分析清楚后，最后在分析难的，或是转化到简单的，或是从简单的解答中，逐渐的增加影响因素，步步向实际问题接近，尽可能的来解决问题。

(六) 讨论

对于食物安全，事实上更关注的是对污染物摄入量的考察，如果人对于某污染物的摄入超过安全标准，就应该预警，不同于食物污染物含量的检查，我们更关注的应该是人摄入某污染物的实际情况。下面提出两种思路：

思路一：

每个人每天会摄入多种食品，某个地区某个年龄段的人在一定时间内摄取主要食物种类相对稳定，无妨设一个人在一段时间内摄取的食物种类为 n 种，且这 n 种食物的摄取量分别为 $x_1, x_2, \ldots, x_n$。设污染物有 m 种，无妨用铅来讨论，设第 i 种食物中铅的含量为 y_i，则一个人一天的摄入量 $z = \sum_{i=1}^{n} x_i y_i$，其中 $x_1 y_1, x_2 y_2, \ldots, x_n y_n$ 相互独立。设 x_i 的密度函数为 $f_i(x)$，y_i 的密度函数为 $g_i(y)$，则 z 的密度函数就是 n 个乘积函数的卷积，即

$$l(z) = (f_1(x)g_1(x)) * (f_2(x)g_2(x)) * \ldots * (f_n(x)g_n(x))$$

当 $n = 1, 2$ 时，比较容易计算它们的分位数。当 n 取值较大时，除了一些特殊的密度函数，如 $x_i y_i$ 都是正态分布或截尾正态分布，我们可以直接从这些分布出发计算出 z 的分布，由此求出分位数，与与国家规定的每天人均污染物摄入量比较，确定是否预警。

思路二：

1．采用自助法(bootstrap)将污染物中小于某值的数据(占总数据量2%)补齐；

2．当食物的摄入量和污染物的含量不匹配，我们假设两次调查是独立进行的，并且认为摄取的食物i的量为总体x_i，污染物含量为y_i，对于给定的一种食物含量x_{ij}，从污染物j含量数据中随机抽取一个数据y_{ij}，做积$z_j = x_{ij}y_{ij}$；

3．采用核估计的办法估计出每个人每天摄入污染物j含量的密度函数$f_j(z)$；

4．利用 $f_j(z)$计算出99.999%的右分位数点；

5．由此分位点与国家规定的每天人均污染物j的摄入量比较，来确定是否预警。

参考文献

[1] 汪荣鑫编著. 数理统计[M]. 西安：西安交通大学出版社，2006

[2] 罗炜，陈冬东，唐英章，李淑娟. 论食品安全暴露评估模拟模型[J]. 食品科技，2007(2):21-24.

[3] 关治，陈景良. 数值计算方法[M]. 清华大学出版社,北京，2002

[4] Duane Hanselman，Bruce Littlefield 编著. 朱仁峰编译. 精通 Matlab 7[M]. 清华大学出版社. 北京，2006

[5] Efron B, Tibshiani R J. An introduction to the bootstrap[M]. New York: Chapman & Hall,1993.

[6] 吴喜之,王兆军.非参数统计方法[M].高等教育出版社,1996.

[7] 张尧庭.金融市场的统计分析[M].广西师范大学出版社,1998.

[8] 张良勇,宋向东,董晓芳，郭照庄.非参数核密度估计[J].佳木斯大学学报,2006,4(24):581-582.

[9] 王茂起,王天竹.2000-2001 年中国食品污染物监测研究[M].卫生研究,2003,32(4):322-326.