

2007 年 A 题

## 建立食品卫生安全保障体系数学模型及 改进模型的若干理论问题

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物加工厂、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人民群众所消费,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。

据初步了解,目前美国和欧盟对公共食品卫生安全实行监控的做法是建立膳食暴露评估数学模型并制成软件,只要将有关的调查或检测数据输入软件,就可以对当时的公共食品卫生安全做出评估。它们所采用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。很可惜美国和欧盟向外提供的软件只是一个黑箱,我们无法断定这个黑箱是否符合我国的实际情况,对他们的数学模型也从加以考证。

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,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。

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物摄入量模型,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。(这个问题只供感兴趣的研究生队选做,也完全可以不做)

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,问题是我们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,即人群食物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,人群食物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，你们的模型采用什么方法提高它的精度？

从以上介绍的大致情况看，其中的确有一些严重的困难和问题，但美国和欧盟已经建立了有关模型并且开始使用，这说明他们已经根据他们的国情将上述困难解决或基本解决，当然也可能仍然含有一些缺陷或存在比较大的误差。

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。例如：1. 污染物含量的分布显然不是正态分布而且很可能是左偏态的，因为污染物含量总是非负的，显然污染物含量越高则概率密度就越小，如何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，或者退一步仅求出其均值。2. 两个不配套的抽样调查数据用什么方法去衔接使用并达到理想的效果，为启发思考，这里抛砖引玉，提出一种思路，例如认为两批抽样调查是完全独立进行的。3. 在调查数据中统计分类标准不相同怎么转化，为启发思考，这里抛砖引玉，提出一种思路，是否可以把调查数据看成是从一些随机变量总体的并集中抽样得到的数据，并设法用它们来估计另一些随机变量（与前者部分相同，部分不同）总体的并集的概率分布函数。4. 当需要利用某些省、市的日常监测数据来估计全国的情况时也面临着两者的概率分布函数可能并不相同的问题，为启发思考，这里抛砖引玉，提出一种思路，是否可以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的并的概率分布函数。这里的难点是我们对比较多的情况都掌握得并不十分清楚，因此你们不妨做出一些必要、合理的假设，并在此基础上进行详细的分析。此外，我们现在没有而且在竞赛的短短几天内也无法使用大规模的调查数据，而建立数学模型又离不开数据，但我们相信参赛的研究生队完全能够创造性地解决这一困难。