DOI:10.13657/j.cnki.gxkxyxb.1983.01.010

全国第二次营养调查
广西、广东体检数据处理程序设计

广西计算中心 刘连芳 张正铀 郑李伶

提 要

用微型机统计分析营养调查中体检数据是计算机用于食品卫生领域的一个新的尝试。广西计算中心和广西防疫站、广东防疫站协作，用国产 BCM—Ⅱ 微型机统计分析了全国第二次营养调查中广西、广东的体检数据，为广西、广东有关部门制定有关政策提供了科学依据。

用计算机处理体检数据，是计算机在卫生防疫领域应用的一个新的尝试。在国内尚没有这方面的经验报导。

体检数据处理就是将全省几千份体检表格按人群(工、农、学、儿童…)、民族(汉、水、瑶…)、调查点（几十个）及性别、年龄的各种组合，统计分析各类数据（如化验结果、测量指标…）、寻找各种情况下各类数据之间的各种内在联系（如18岁以下青少年的年龄与身高、体重的关系，学龄前儿童某些摄入量与相应临床、化验指标的关系…）、为了解各种人群、各个地区、各民族、各种年令人的身体状况，为医务工作者制定临床诊断疾病标准提供可靠依据，为防疫部门开展防疫工作，为提高人民的健康水平，给政府有关部门制定有关政策，提供了科学依据。

体检数据按年龄可分为两大部分。一部分为7岁以前儿童的，需做十类400余种统计分析；另一部分为7岁以上（含7岁）人员的，需做十类200余种统计分析；两部分混合处理项目约110种。关系复杂，工作量大，人工很难在短期内完成。

我们与广西防疫站和广东防疫站共同协作，用国产 BCM－Ⅱ 微型机处理了全国第二次营养调查广西、广东体检数据。每省、区原始数据均分别为20万左右。工作自1982年12月28日开始历时一个多月。

BCM－Ⅱ型机内存64K，可用两个8吋双面单密度软磁盘。

为了尽量发挥计算机的效力，减少人工劳动（如编码、键盘输入等）；为在内存容量的允许下，尽可能提高处理速度，我们先对数据编码、输入和体检采取了必要措施，并对数据进行了预处理。整个软件共分三部分计二十五个独立的程序。

一、编码及键盘输入

由于体检表格是为防疫部门人工处理而设计的，所以有些数据的安排对于用计算机处理

本文1983年4月12日收到

来讲不够合理。比如，化验结果所对应的临床印象诸项的顺序有倒排现象。为减少人工工作，诸如此类问题留给计算机处理，尽量不变动原体检表格。

人工预先只需作两方面的工作：

1．性别编码。

2．统计体征及临床印象中“+”（不正常）、“？”号（怀疑）的个数及其在表中的位置（原表格中大多数项均有序号）。

做第二项工作虽然稍微增加了点工作量，但可减少大量键盘输入。每个学令前儿童的体征和临床印象共54项，7岁以上人员（含7岁）是49项，若全部输入，需分别打108和96个字符。但是，考虑到绝大多数人的绝大多数项均为“－”号（正常），所以采用输入特殊信息、隐含一般信息的方法，即只输入每个人“？”号、“+”号个数及其在表格中所在位置。全为“－”号的，则用两个0表示（一个代表“+”号零个，一个代表“？”号零个）。这样，即使在最坏的情况下，也只需用十几个字符（包括分隔符）即可表示四十、五十项数据。由于大多数人的相应项中只需输入两个零，所以平均每个人只需7、8个字符。体征、临床印象两部分数据的输入，一共可减少90%以上。

另外，凡单项缺值则用－9代替，若某一类缺值，例如化验结果全部没有，则用一个标识符（定义为33333）表示，既减少输入数据量也便于以后的处理。

为方便计算机检查数据错误，每个人的全部数据输入结束后打入一个界符（定义为99999）。

二、程序设计

1．**数据检查**。数据检查程序可使计算机检查出漏打、多打、超出正常值范围的数据及标点错误，指出错误所在处。

对全部数据进行检查用机约1小时，比人工检查准确且大大提高速度，尽量避免由于数据错误导致以后处理上的失败。

2．**数据预处理**。由于内存容量的限制，不可能一次做完所有统计分析，要分若干次统计则要多次检索数据。比如，广西参加体检总人数为2910，而学令前儿童仅为578，学令前儿童的体检项目大多数与其他人不同。所做统计分析也不同，大部分必须单独处理。多次从2910人的数据中检索出578人的数据，需要读取80%的无用数据及进行大量“资料是否为所需”的判断。对于7—18岁和18岁以上（含18岁）情况也如此。并且，对于同一年令组的数据，每次处理只需其中一部分，或是测量指标、或是体征、或是化验结果等等，不必每次取出全部数据。鉴于以上情况，我们在进行统计计算之前对数据进行了预处理。将所有数据按年令分为三组：7岁以下；大于等于7岁、小于18岁；大于等于18岁。每个年令组又按数据类别分成五个固定结构文件：①、代码（包括点号、人群号、户号及性别、年令），②、测量指标（身高、体重、头围…）；③、体征；④、化验结果；⑤、临床印象。这样，在进行统计分析时只要打开所需年令组中相应的二、三个文件即可，大大减少不必要的操作，提高速度。

每个年令组内的五个文件之间用各记录的相对位置来联系。由于各文件内数据类型相同，长度相近，据最大数据长度定义的记录长度可以尽量减少磁盘空间的浪费。

同时，在对数据进行预处理时，将输入数据中的标识符转换成了体检表格中的数据。这

样虽然多占用了磁盘空间，但在检索时可以大大减少判断，提高速度，以磁盘空间换取了运算时间。

数据顺序在预处理中也做了必要的调整。

数据预处理共生成15个文件。使用这些文件时，若其中一个产生意外，只需解决这一个文件的问题，对其他文件没有影响。这样，既便于排除数据文件“故障”，又减少工作量。

全部数据处理共用20小时。

数据处理程序框图见框图1

3．统计分析。全部统计分析共用23个程序。除回归程序外，其他程序的基本共同点是必须作两类转换。一是将输入的代码中点号、人群号转换为计算机统计分析用的序号，输出时再进行逆转换。例如，对广西某三十个点7岁以上人员（含7岁）各体征进行统计时，南宁的散居点人群号为9，点号为1，在程序内序号为15；而同一点在同一程序内进行人群、民族情况统计时，序号又分别为5（城市），10（散居户）。二是将性别、年令与程序内的组号之间进行转换。有了序号、组号，即可方便地时各类体检数据进行统计分析。在各程序中统计项目和要求均不同，则序号、组号也随之不同。输出结果时，为使用户阅读直观、方便，在程序内均根据用户工作习惯进行了逆转换。

运行23个统计计算程序共用机40小时，比人工统计大大提高了速度。正式运算前，全部程序经过了正确性验证，所以提供的结果准确可靠。

程序实例见框图2

结 束 语

统计中所用营养摄入量文件系陈振坤、张正铀所做“膳食调查数据处理”的一个结果文件。

统计中进行 X^2 检验和求 S、求参考正常值的方法及公式由广西防疫站陈正清医师提供。陈医师、杨以祯医师、卫庆远医师及广东防疫站魏文强医师和邓峰医师参加了数据输入和结果分析。

用计算机统计所花经费仅为防疫部门原计划用于该项目经费的10%，效率提高几十倍。

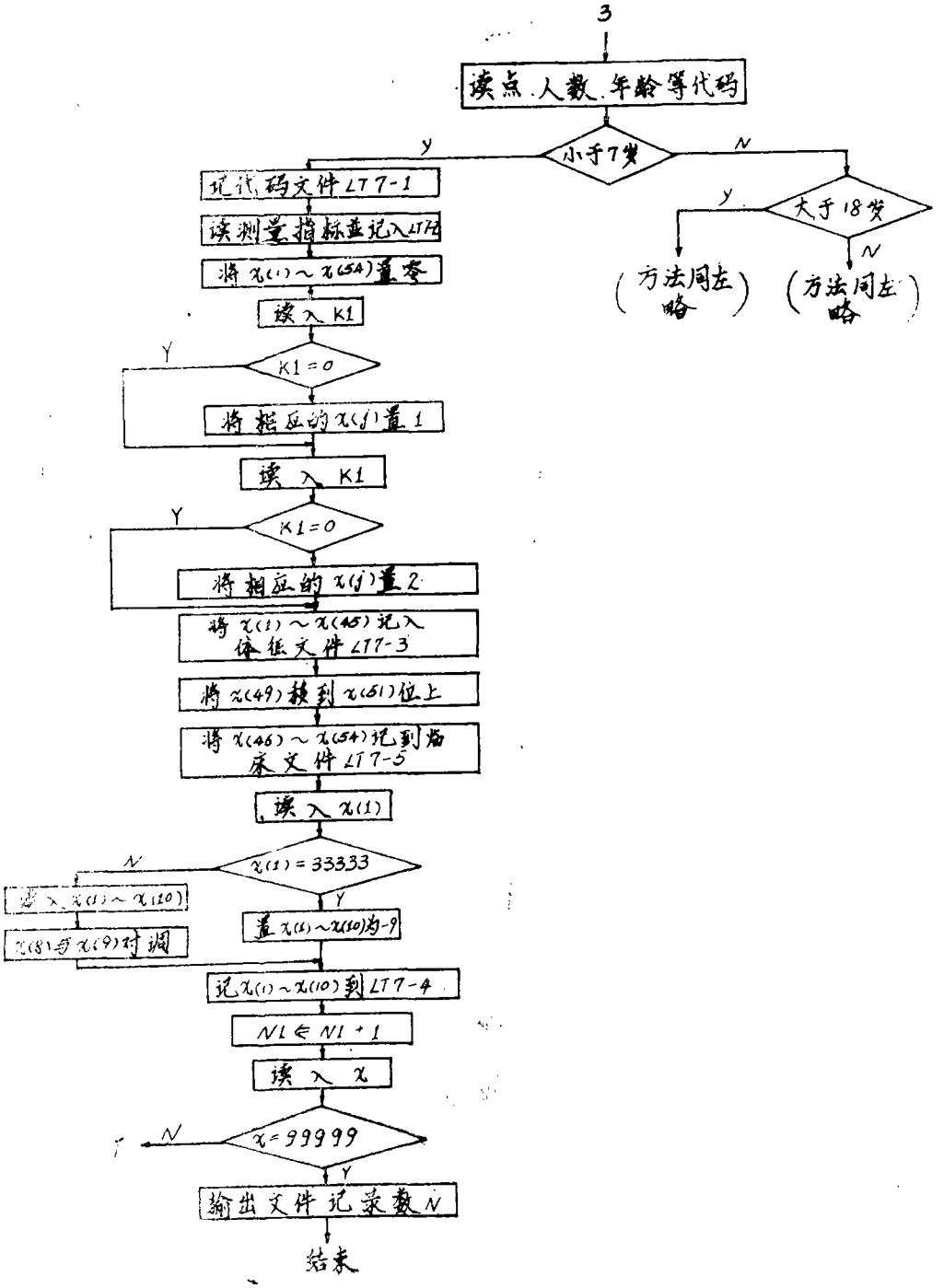
3
读点、人数、年龄等代码
小于7岁
y
N
大于18岁
y
N
(方法同左略)
(方法同左略)
记代码文件 LTT-1
读测量指标并记入 LTT-2
将 X(1)～X(54)置零
读入 K1
K1=0
Y
将相应的 X(J)置1
读入 K1
K1=0
Y
将相应的 X(J)置2
将 X(1)～X(45)记入体征文件 LTT-3
将 X(49)移到 X(51)位上
将 X(46)～X(54)记到指标文件 LTT-5
读入 X(1)
X(1)=33333
N
Y
读入 X(1)～X(10)
X(8)与 X(9)对调
置 X(1)～X(10)为-9
记 X(1)～X(10)到 LTT-4
N1←N1+1
读入 X
X=99999
N
Y
输出文件记录数 N
结束

图1

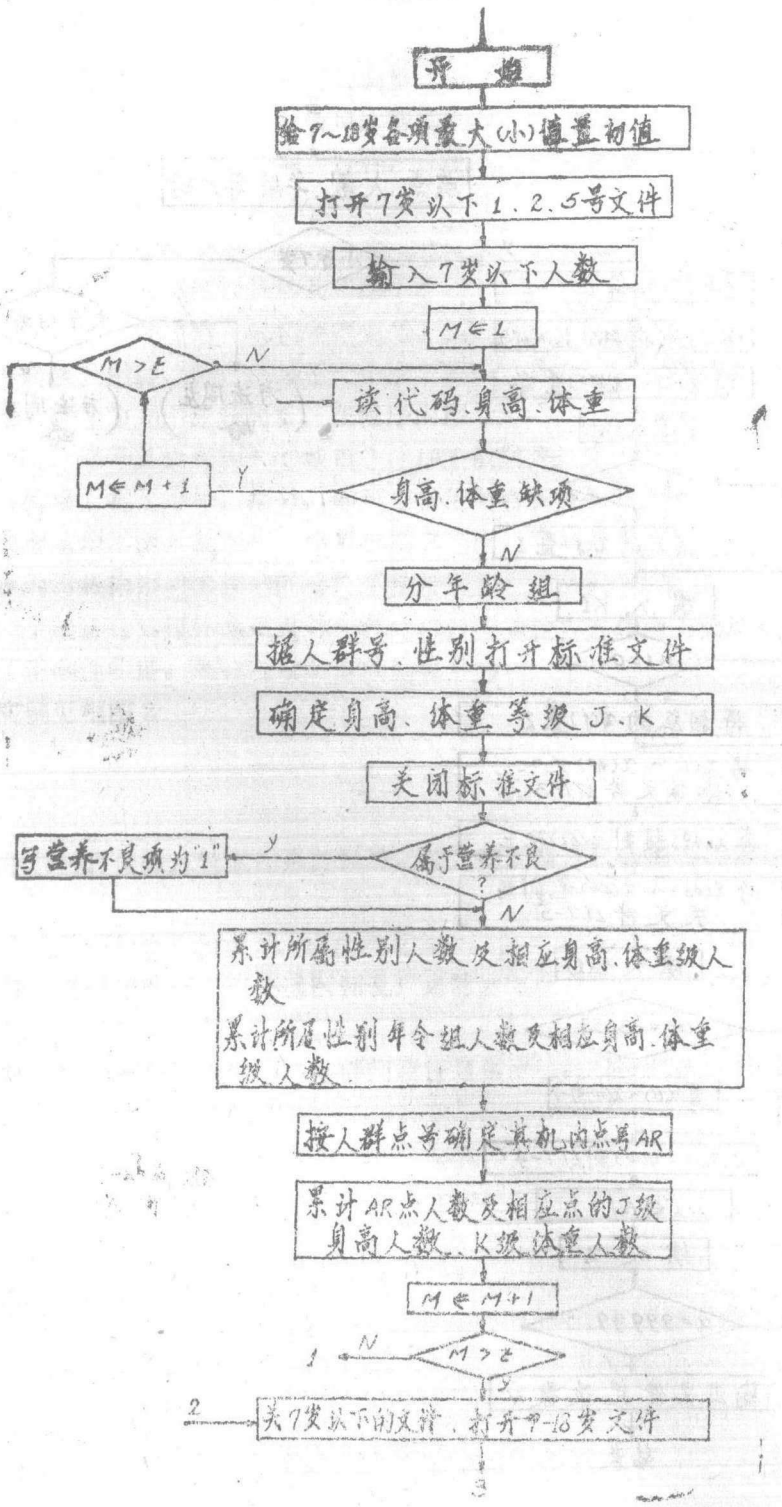
开 始
给7~18岁各项最大(小)值置初值
打开7岁以下1、2、5号文件
输入7岁以下人数
M←1
M>E
N
读代码、身高、体重
M←M+1
Y
身高、体重缺项
N
分年龄组
据人群号、性别打开标准文件
确定身高、体重等级J、K
关闭标准文件
写营养不良项为"1"
y
属于营养不良?
N
累计所属性别人数及相应身高、体重级人数
累计所属性别年令组人数及相应身高、体重级人数
按人群点号确定其机内点号AR
累计AR点人数及相应点的J级身高人数、K级体重人数
M←M+1
1
N
M>E
Y
2
关7岁以下的文件,打开7-18岁文件
3

图 2

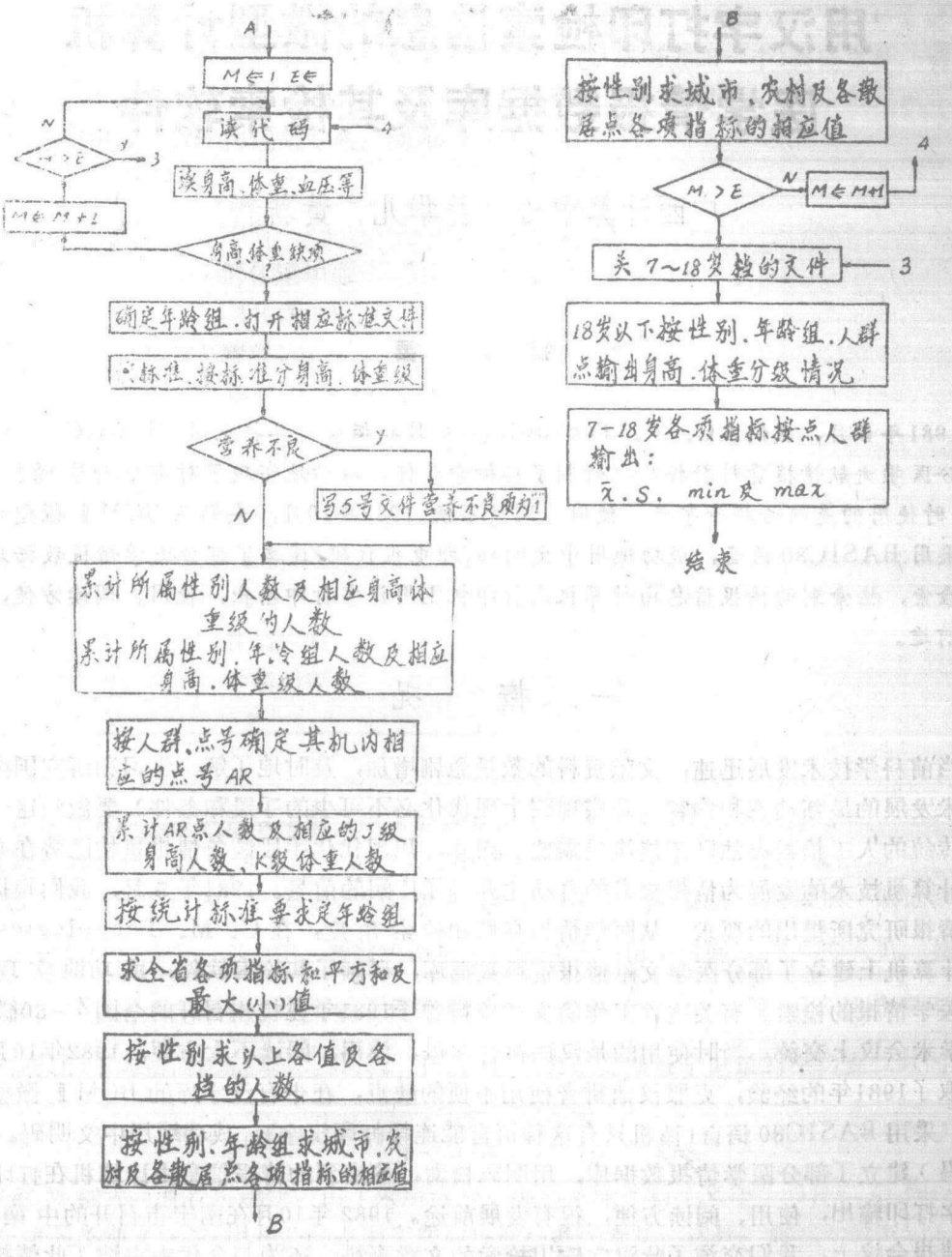
A
M←1 E←
读代码
4
N
M>E
Y
3
读身高、体重、血压等
M←M+1
身高体重缺项?
确定年龄组，打开相应标准文件
标准、按标准分身高、体重级
营养不良
Y
写5号文件营养不良项为1
N
累计所属性别人数及相应身高体重级的人数
累计所属性别、年龄组人数及相应身高、体重级人数
按人群、点号确定其机内相应的点号AR
累计AR点人数及相应的J级身高人数、K级体重人数
按统计标准要求定年龄组
求全省各项指标和平方和及最大(小)值
按性别求以上各值及各档的人数
按性别、年龄组求城市、农村及各散居点各项指标的相应值
B
B
按性别求城市、农村及各散居点各项指标的相应值
M>E
N
M←M+1
4
关7~18岁档的文件
3
18岁以下按性别、年龄组、人群点输出身高、体重分级情况
7-18岁各项指标按点人群输出：
x̄、S、min及max
结束

图3