数学无处不在 – 血液检测的数学模型和理论（二）

胡晓东
中国科学院数学与系统科学研究院

今天我继续给大家介绍与血液检测相关的数学模型和理论：如何应用数学的方法，用最少的成本（检验次数和时间），在大量的血液样本中准确地检验出哪些是阴性的（好的），哪些是阳性的（坏的）。

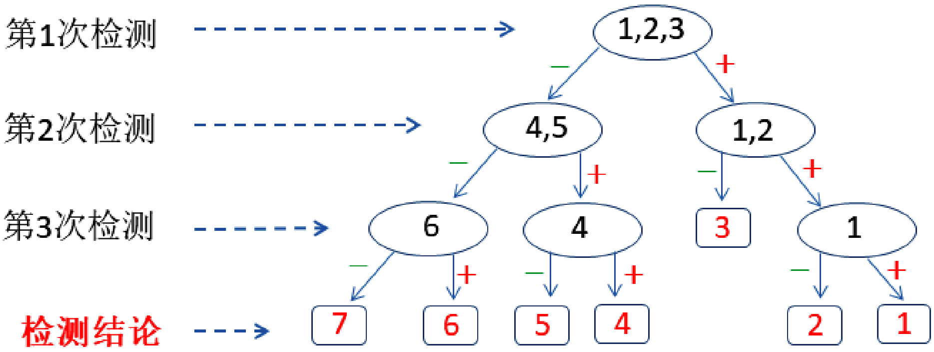
上一次我讲到，1943 年 R. Dorfman [1]提出了一个组合检测法：将若干个样本放在一起进行检测。如果检测结果显示是阴性的，那么这个组里的所有样本都是阴性的；如果检测结果显示是阳性的，那么再将这个组里的样本逐一检测。1960 年 P. Ungar [2]对这个问题进行了研究。他考虑的是一种**概率模型**：假设所需检测的 **N** 个样本有 ***p*N** 个是阳性的，其中 $0 \le p < 1$；每一个样本可以有很多备份，每组可以包含任意多的血液样本（备份）。在这些假设下，他证明：当 $p < 0.382$ 时，组合检测法平均需要的检测次数比逐一检测法需要的检测次数少。

应用实践和理论分析都表明，当大量的待检测样本中坏的样本不是很多时，比如，高考考生体检验血，或者新兵入伍复查体检验血，组合检测法比逐一检测法需要的检测次数要少。

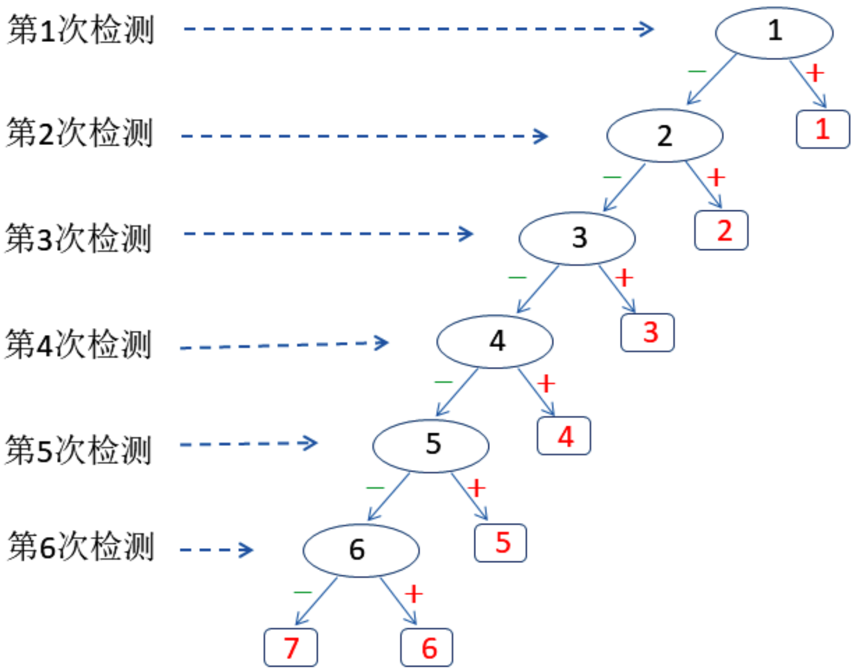

今天我来介绍**组合模型**：假设事先已经知道所给的 **N** 个样本中有 d 个是阳性的，其中 $0 < d < N$。（未经检测就知道有 d 个是阳性的，这个假设不合理，是吧？我将在下一次介绍当这个假设不满足时，如何设计组合检测法）。下面先考虑最简单的情形，看看组合检测法好呢？还是逐一检测法好呢？

当 $d=1$ 时（只有 **1** 个样本是坏的），可以二分技巧快速检测到这个坏的样本，亦即将 **N** 个样本尽量平均分成两组 A 和 B，先检测 A 组；

- 如果 A 组的检测结果显示是阳性（意味着 B 组中的样本都是好的），那么再将 A 组中的样本尽量平均分成两组 A1 和 A2，并检测 A1 组 … … 。
- 如果 A 组的检测结果显示是阴性（意味着 A 组中的样本都是好的），那么将 B 组中的样本尽量平均分成两组 B1 和 B2，并检测 B1 组 … … 。



上面的图示给出了用组合检测法对 **N=7** 个样本的检测过程，检测开始前已知只有 **1** 个是坏样本，但是不知道哪一个样本是坏的。可以看出：当样本❸是坏样本时，用 **2** 次检测就可以找到它；但是其他情况都需要 **3** 次检测。经过简单的计算可知：基于二分技巧的组合检测法用 $\lceil \log_2 N \rceil$ 次检测一定可以找到那个坏的样本。



上面的图示给出了用逐一检测法对 **N=7** 个样本的检测过程。可以看出：当样本❶是坏样本时，用 **1** 次检测就可以找到它；但是当样本❻或者❼是坏样本时，则需要 **6** 次检测。显然，逐一检测法需要 **N-1** 次才能确保找到那个坏的样本。

另外，当 d = N-1 时（只有 1 个样本是好的），此时使用组合检测法是没有任何意义的，因为将任意两个样本放到一起进行检测，（不用检测就可以推断出）结果一定是阳性的，所以逐一检测法是最好的。

上面是两个最极端的情形：只有一个样本是坏的，或者只有一个样本是好的。前一种情形，组合检测法最好，后一种情形，逐一检测法最好。下面我们来考虑非极端的情形。比如，N=9，d=2，采用组合检测法 7 次检测就可以了（建议你自己试一试），比逐一检测法用的检测次数少一次。而当 N=9，d=3，你会发现采用组合检测法需要 8 次检测才可以，与逐一检测法用的检测次数是一样多。

经过上面的初步分析，我们自然就会提出如下的数学问题：

> 什么时候组合检测法比逐一检测法用的检测次数少？亦即，给定 N 个样本，并且已知其中有 d 个样本是坏的，当 d 不超过多少时组合检测法用的次数少于 N-1?

1971 年 F. K. Hwang（黄光明）[3]研究了上述问题。他用 M(N, d) 表示最优检测法所需次数。显然 M(N, d) ≤ N-1，当不等式取等号时，就表示逐一检测法是最优的。他证明：当 2d + 1≥N 时，M(N, d) = N-1，亦即，若已知有 1/2 以上的样本是坏的，则逐一检测法是最好的。

1981 年 F. K. Hwang 与合作者[4]通过进一步的研究，证明：当 5d+1≥2N 时，M(N, d) = N-1，亦即，若已如有 2/5 以上的样本是坏的，则逐一检测法就是最好的。同时，他们提出了如下猜想：

> **Hu-Hwang-Wang 猜想：** 当 d ≥ N/3 时，M(N, d) = N-1。

上述猜想成立意味着：若已知有 1/3 以上的样本是坏的，则逐一检测法是最好的。

1982 年 D.-Z. Du（堵丁柱）与 F. K. Hwang [5]合作对上述猜想进行了研究。他们证明：当 21d ≥ 8N 时，M(N, d) = N-1，亦即，若已知有 8/21 以上的样本是坏的，则逐一检测法就是最好的。尽管 8/21 已经非常接近 1/3=7/21，这个结果也已发表近四十年了，但是至今还未出现更好的结果。Hu-Hwang-Wang 猜想 已成为了组合检测/搜索研究方向的一个著名难题。大家有没有兴趣挑战一下？

下一次我将给大家介绍一下，当我们得到了 **N** 个（血液）样本，但不知道关于它们的任何信息：既不知道好的样本多还是坏的样本多，更不知道究竟有多少个是坏的样本，我们又该如何设计组合检测方法呢？

参考文献

[1] R. Dorfman, The Detection of Defective Members of Large Populations, The Annals of Mathematical Statistics, 14 (4)(1943), 436-440.

[2] P. Ungar, The Cutoff Point for Group Testing, Communications on Pure and Applied Mathematics, XIII (1960), 49-54.

[3] F. K. Hwang, A minimax procedure on group testing problems, Tamkang Journal Mathematics, 2 (1971), 39-44.

[4] M. C. Hu, F. K. Hwang and J. K. Wang, A boundary problem for group testing, SIAM Journal on Algorithms and Discrete Methods, 2 (2) (1981), 81-87.

[5] D.-Z. Du and F. K. Hwang, Minimizing a combinatorial function, SIAM Journal on Algorithms and Discrete Methods, 3 (1982), 523-528.