Mathematics is Everywhere

数学无处不在——核酸混合检测法

胡旭东

一、题 记

2019 年 11 月 26 日联合国教科文组织在第四十届大会上批准并宣布，每年的 3 月 14 日为“国际数学日”，设立该纪念日的目的是为了让越来越多的人去感受“数学在生活中的美丽与重要”。该项目是由国际数学联盟发起和领导的，并得到了来自世界各地众多的国际和地区组织的支持。

第一个国际数学日是 2020 年 3 月 14 日，主题是“数学无处不在”（Mathematics is Everywhere），强调数学在人们日常生活的几乎每个领域都发挥着至关重要的作用：从自然模式到气候科学，从医学成像到搜索引擎，从运输网络到 AI 的优化，等等。国际数学日扩大了以前在 3 月 14 日庆祝“圆周率日”的范围，涵盖了整个数学领域并延伸到了整个世界的方方面面。然而，非常遗憾的是，2020 年初开始，新冠疫情在全球范围大流行，原计划于 2020 年 3 月 13 日举行的联合国教科文组织国际官方发布会及相关活动被迫取消。

在新冠肺炎开始在全球迅速蔓延的初期，患病人数持续暴增。确诊患者的一个重要方法就是进行核酸检测。然而，大量的核酸检测不仅会耗费大量的人力、物力和财力，而且还需要花费很多的时间，给各国政府和人民在经济和生活中都造成了极大的负担和影响。

2020 年 4 月 6 日斯坦福大学医学院的研究团队发表了一篇回顾性研究简报[1]。提出了开展核酸检测的一个新思路：对于常规呼吸道病毒检测结果为阴性的患者，采集鼻咽拭子和支气管肺泡灌洗样本，将样本混合为“样本池”后一并检测，可

[1] C. A. Hogan, M. K. Sahoo, B. A. Pinsky, Sample pooling as a strategy to detect community transmission of SARS-CoV-2, The Journal of the American Medical Association, 323(19) 2020,1967-1969.

能有助于改善检测效率，促进对潜在社区传播的早发现。这一策略已用于某些传染性疾病（例如沙眼）的社区监测，但尚未用于美国对新冠病毒感染的早期筛查。研究团队共筛查了 292 份样本池，总共 2888 份单独样本，经确认的新冠病毒阳性样本 2 例。阳性结果来自 2020 年 2 月 21–23 日采集的鼻咽拭子，2 份阳性样本中均检出了新冠病毒包膜和病毒的 RNA 依赖性 RNA 聚合酶（RdRp）基因。研究团队指出，这些筛查结果支持大流行初期旧金山湾区的新冠病毒感染率不高，混合样本池的阳性检出时间也与疾控中心确认当地前 3 名感染者的时间重叠。

2020 年 6 月 5 日中央电视台在焦点访谈节目中报道：从 2020 年 5 月 14 日开始的 19 天的时间里，武汉全市没有进行过新冠病毒核酸检测的常住居民和暂住居民近 990 万人接受核酸检测，加上此前已做过检测的人员，武汉累计共有 1090.9 万人完成核酸检测，基本做到人员全覆盖。检测结果没有发现确诊病例，检出无症状感染者 300 名，检出率为万分之 0.303。

那么近千万人的核酸检测是怎么有效有序组织的？那么大规模的检测靠什么来保证准确性？除了原有的 23 家检测机构以外，武汉市动员了全市 40 家医疗机构和疾控中心参与其中，第三方检测机构检测人员由 419 人增加至 1451 人，检测设备从 215 台套增加到 701 台套。

报道中还特别地提到，为了提高核酸检测的效率，此次检测除了单人单管单样本之外，有 1/4 的采样样本采取了“混和检测法”，最多把 5 份样本混在一个检测管里进行检测，混合样本的检测结果当中一旦检出异常，就会对这一份混合样本中每一个个体样本再进行单独检测。为了在提高效率的同时确保准确度，还设置了必要的抽检复测。对各家检测机构又抽取了 35961 份样本进行复测，结果和初测结果完全吻合，表明混合检测法在可以显著地提高检测效率的同时还能够保证检测的准确性。

2021 年和 2022 年国际数学日的主题分别是“数学让世界更美好”（Mathematics for a Better World）和“万物皆数”（Mathematics Unites）。然而，由于当时新冠病毒在全球范围还在持续肆虐，相关活动不得不取消或者改为线上的形式进行。这里值得一提的是，在 2021 年国际数学日的官网上，列举并介绍了数学的 23 个应用领域，其中就包括“混合检测”（Group Testing），见图 1。

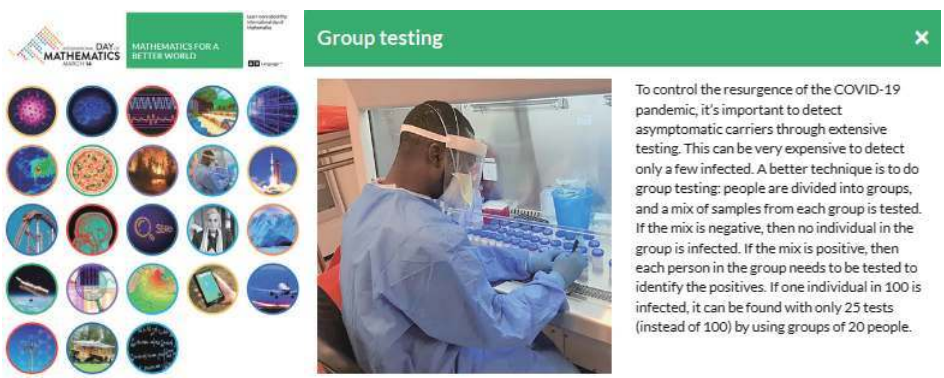

图 1. 2021 年国际数学日的官网首页 https:betterworld.idm314.org

此外，2021 年北京市高考数学试卷的第 18 题就是关于混合检测法的，有兴趣的读者可以试着解答一下。

在核酸检测中，“k 合 1”混采核酸检测是指：先将 k 个人的样本混合在一起进行 1 次检测，如果这 k 个人都没有感染新冠病毒，则检测结果为阴性，得到每人的检测结果都为阴性，检测结束；如果这 k 个人中有人感染新冠病毒，则检测结果为阳性，此时需对每人再进行 1 次检测，得到每人的检测结果，检测结束。

现对 100 人进行核酸检测，假设其中只有 2 人感染新冠病毒，并假设每次检测结果准确。

(I) 将这 100 人随机分成 10 组，每组 10 人，且对每组都采用“10 合 1”混采核酸检测。

(i) 如果感染新冠病毒的 2 人在同一组，求检测的总次数。

(ii) 已知感染新冠病毒的 2 人分在同一组的概率为 1/11。设 X 是检测的总次数，求 X 的分布列与数学期望 $E(X)$。

(II) 将这 100 人随机分成 20 组，每组 5 人，且对每组都采用“5 合 1”混采核酸检测。设 Y 是检测的总次数。试判断数学期望 $E(Y)$ 与（I）中 $E(X)$ 的大小。（结论不要求证明）

自新冠肺炎病毒大流行三年多以来，国内外对新冠疫情的防控实践和理论研究都表明，及时、快速和准确地开展核酸检测是减缓和阻断新冠病毒传播的一个非常重要的环节，而混合检测法是提高病毒感染筛查的一个十分有效的技术手段。本文将介绍混合检测法的历史和相关的数学模型及问题的研究进展，对相关内容特别感兴趣的读者可以参阅专著[2]。

二、概率模型

在第二次世界大战后期，美国政府征集大量的年轻人参战。在征兵过程中需要体检，其中一个环节就是验血，用来筛查出应征参军的报名者中携带淋病病毒的年青人。通常的筛查方法是采集血样以后，对它们一个一个地进行检测，不妨称之为“逐一检测法”。1943 年，美国政治经济学家、管理科学学会第十二任主席多夫曼（Robert Dorfman）发表了一篇文章[3]，为美国公共健康服务和筛选服务系统设计了一个“混合检测法”，后来被称为“多夫曼

[2] D.-Z. Du and F. K. Hwang, Combinatorial Group Testing and its Applications, 2nd edition, World Scientific Publishing Co. Pte. Ltd., Singapore, 2000.

[3] R. Dorfman, The detection of defective members of large populations, The Annals of Mathematical Statistics, 14 (4) (1943), 436-440.

筛查”（Dorfman Screening）：

- 首先，将所要检测的 N 个样本中的每一个样本提取一小部分，剩余部分待用（类似于现在兴奋剂检测采用的 A 瓶和 B 瓶方法）；
- 然后，将 N 个提取出来的样本分成 $N/5$ 组，每组 5 个样本充分混合，得到了 $N/5$ 个样本组；
- 最后，将 $N/5$ 个样本组，一个组、一个组地进行检测。每一个样本组的检测结果可能出现以下两种情况：

（1）检测结果是阴性的，说明该样本组是纯净的（Pure），其中的每一个样本都是阴性的。这样我们检测 1 次就确认了 5 个样本都是阴性的，从而比逐一检测法节省了 4 次检测。

（2）检测结果是阳性的，说明该样本组被污染了，其中至少有一个样本是阳性的，但是并不清楚哪一个或者哪几个是阳性的；因而需要对该组中 5 个样本的剩余样本，进行逐一检测以便确认每个样本是阴性的还是阳性的。这样总共通过 6 次检测才可以确认 5 个样本中哪些样本是阴性的哪些是阳性的，比逐一检测法多用了 1 次。

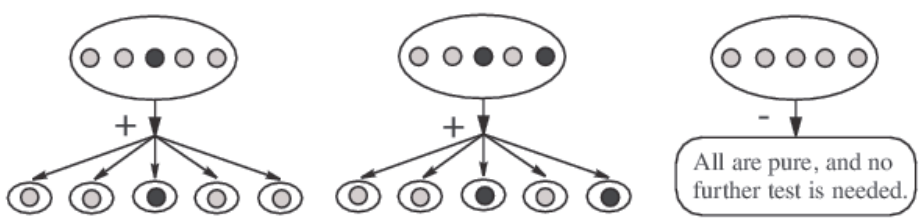

图 2. 多夫曼筛查示例

不难看出，如果在所有样本中阳性的比较少，那么混合检测法会减少大量的检测次数；而这正是当年征兵进行血液检测时的情形（现在每年高考进行体检时也类似）。可是如果在所有样本中阳性的比较多，那么混合检测法很有可能不仅不会减少检测的次数，还会增加检测的次数。多夫曼由此提出如下两个非常自然的问题：

(I) 混合检测法是否比逐一检测法用的检测次数少？如果是，可以减少多少次呢？

(II) 每一组中混合多少个样本是最优的呢？

1960 年，昂加尔（Peter Ungar）对问题 (I) 进行了研究，他称其为“混合检测问题”（Group Testing Problem）。他假设所需检测的 N 个样本有 pN 个是阳性的，其中 $0 \leqslant p < 1$；每组可以包含任意多个样本（不再限定最多 5 个样本），而且每一个样本可以重复使用（相当于同一个样本有很多个备份）。在这

些理想的假设下，他证明[4]：当 $0 \leqslant p < (3-\sqrt{5})/2 \approx 0.38$ 时，必定存在一种混合检测法，其需要的检测次数期望值少于 N。

关于问题 (II)，若令每组混合 n 个样本，每个样本为阳性的概率为 p，则阴性的概率为 $q = 1-p$，样本组中每一个样本都是阴性的概率为 q^n，样本组中至少有一个样本是阳性的概率为 $1-q^n$，可计算出混合检测法需要检测次数的期望值为 $\mathrm{E}(N, q, n) = N[q^n + (n+1)(1-q^n)]/n$。给定 N 和 q，即可求出最优的 n 使得 $\mathrm{E}(N, q, n)$ 达到最小值（见图 3）。

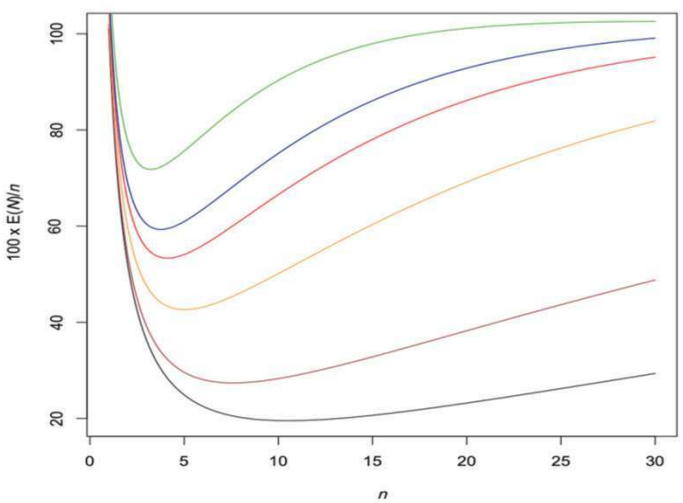

图 3. 固定感染率 p，函数 $100 \times \mathrm{E}(N)/n$ 表示若每组含有 n 个样本，需要检测的平均次数。其中 1%（黑线），2%（棕线），5%（橙线），8%（红线），10%（蓝线）和 15%（绿线）

多夫曼和昂加尔等人对大规模血液样本检测问题的研究激发了很多后续的研究，各种各样的检测和搜索问题的数学模型和方法也先后被提出来，并逐渐形成了一个数学研究方向：组合搜索[5]。

三、组合模型

这一节我们讨论混合检测法的组合模型：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 个样本是阳性的，这不符合实际情况。我们将在下一节讨论当该假设不满足时，如何设计混合检测法）。下面先考虑最简单的情形，看看混合检测法好呢？还是逐一检测法好呢？

[4] P. Ungar, The Cutoff point for group testing, Communications on Pure and Applied Mathematics, XIII (1960), 49-54. P. Ungar, The Cutoff point for group testing, Communications on Pure and Applied Mathematics, XIII (1960), 49-54.

[5] M. Aigner, Combinatorial Search, Stuttgart: B.G. Teubner; Chichester, England; New York: J. Wiley, 1988.