数学无处不在 – 血液检测的数学模型和理论（五）

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：如何运用数学的方法，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）。

前四次我先后讲了四个模型，**概率模型**：假设事先已经知道所需检测的 **N** 个样本中有 pN 个是阳性的，其中 $0 \le p < 1$；**组合模型**：假设事先已经知道所需检测的 **N** 个样本中有 d 个是阳性的，其中 $0 < d < N$；**竞争模型**：事先既不知道 p 也不知道 d；**容错模型**：检测结果可能出现错误。同时介绍了相应的四个理论结果。

上述四个模型都属于**序贯方法**：每次检测哪些样本组，通常要依赖以前的检测及其结果。一个序贯方法如果需要进行 n 次检测才能检测出所有的坏样本，而每次检测需要 t 时间的话，整个检测过程通常需要 $n \times t$ 时间。如果想缩短时间，那么除了尽可能地减少检测次数，还有一个方法就是在 t 时间内同时（并行地）检测若干（而不是一个）样本组。

今天我要给大家介绍一种**非序贯方法**：在检测开始前，需要把将要检测哪些组完全确定下来，换言之，检测哪个组并不依赖于其他检测组的结果。不难发现，尽管非序贯方法用的检测次数不会少于序贯方法，但是完成全部测试所需用的时间段长度会小于序贯方法，因为非序贯方法允许同一时间段并行地进行检测（只要有足够多的检测设备；比如，如题图所示，有 **4** 台设备）。逐一检测法可以看作是非序贯检测法，因为所有的检测都可以在时间 t 内同时完成（但是至少需要 **N-1** 台检测设备）。

首先我想给大家回顾一下《环球时报》4 月 5 日的一则报道：“…疫情在全球蔓延后，比尔盖茨更是耗费数十亿美元投资研发新冠疫苗。据《国会山报》4 月 3 日报道，比尔盖茨的基金会目前正在投资建造工厂，这些工厂将生产七种有前途的冠状病毒疫苗。比尔盖茨对《每日秀》的主持人“崔娃”表示，虽然他最后只会从这七种疫苗中选择一到两种，但他也会出钱让七个工厂同时运作，这样就不会因为一个个排除而浪费时间 …”可以看出，比尔盖茨采用的就是非序贯方法。

下面言归正传，我们先看组合模型的一个简单例子：在待检测的 **4** 个样本中，已知仅有一个是坏样本。若采用二分策略的组合检测（序贯方法），用 **2** 次检测就可以找出那个坏的样本；如果做一次检测的时间是 ***t***，那么完成两次检测的时间就是 **2*t***。能不能用 ***t*** 时间就检测出那个坏样本呢？

样 本	①	②	③	④	坏样本	❶	❷	❸	❹
检测 I	✓		✓		结果 I	+	−	+	−
检测 II		✓	✓		结果 II	−	+	+	−

上面的图表给出了一个非序贯的组合检测方案：检测 I 包含样本①和③，检测 II 包含样本②和③。同时进行检测 I 和检测 II。若得到检测结果 I 和结果 II 是“+”和“−”，则样本①是坏样本；若得到检测结果 I 和结果 II 是“+”和“+”，则样本③是坏样本；等等。这个方法就可在 ***t*** 时间里同时完成两次检测，而且一定能找出那个坏样本。一般的，若已知在待检测的 **N** 个样本中，有惟一一个是坏样本，则可以在 ***t*** 时间内同时完成 $\lceil \log_2 N \rceil$ 次检测，并找出那个坏样本。请大家自己设计这样的一个非序贯的组合检测方案吧，你一定能行！

最后，我来给大家介绍一个**猜年龄游戏**。实际上，它是前一次我讲的**乌拉姆猜数游戏**的一种特殊情况（我在上一次已经说明：猜数游戏实际上是血液检测问题的特例）。考虑到她的年龄不超过 **60** 岁，且 $60<2^6$，因而我可以采用基于二分策略的序贯方法，向她提问最多不超过 **6** 个问题，一定可以猜出她的准确年龄。而在这个游戏中，我给出了 **6** 张表，每张表相当于一个问题：她的年龄是否出现在这张表中？只不过这次，这 **6** 个问题是一次给出的（非序贯方法），而不是像采用二分策略（序贯方法），一个问题问完了以后，根据得到的回答，再问下一个问题。

猜年龄游戏：我知道她今年不会超过**60**岁，但不知道她准确的年龄。我准备了下面六张表。如果她能告诉我，她的年龄都出现在了哪几张表中，我就能猜出她的年龄☺

3	5	7	9	1	11
13	15	17	19	21	23
25	27	29	31	33	35
37	39	41	43	45	47
49	51	53	55	57	59

3	6	7	11	2	10
14	15	18	19	22	23
26	27	30	31	34	35
38	39	42	43	46	47
50	51	54	55	58	59

5	6	7	13	4	12
14	15	20	21	22	23
28	29	30	31	36	37
38	39	44	45	46	47
52	53	54	55	60	13

9	10	11	12	8	13
14	15	24	25	26	27
28	29	30	31	40	41
42	43	44	45	46	47
56	57	58	59	60	13

17	18	19	20	16	21
22	23	24	25	26	27
28	29	30	31	48	49
50	51	52	53	54	55
56	57	58	59	60	31

33	34	35	36	32	37
38	39	40	41	42	43
44	45	46	47	48	49
50	51	52	53	54	55
56	57	58	59	60	46

比如，当她告诉我，她的年龄出现在第①、③、④、⑥张表中的时候，我就猜出了她的年龄是 **45** 岁。我是如何做到的呢？首先，我观察到 **6** 张表的每张表第 1 行第 5 个数字分别是 **1**，**2**，**4**，**8**，**16**，**32**。然后，我将第①、③、④、⑥张表中对应的数字相加得到 **1+4+8+32=45**，这就是我猜出的年龄。大家不妨验证一下，**45** 的确出现而且仅出现在第①、③、④、⑥张表中。

我为什么能利用这 **6** 张表猜出她的年龄呢？注意，每一个自然数都可以用二进制惟一表示出来，比如，$2^0+2^2+2^3+2^5=45$。而这 **6** 张表的每一张表第 1 行第 5 个数字分别是 2^0，2^1，2^2，2^3，2^4，2^5，它们分别都仅出现在一张表中。因而，可以将不超过 **60** 的任何一个自然数，根据其二进制惟一表示，填入相应的表格中。显然，将 **1**，**2**，**4**，**8**，**16**，**32** 分别放在 **6** 张表的每张表第 1 行第 5 列的位置，只是为了计算方便，实际上可以将它们分别放到每张表的任何位置上。建议大家如法炮制，制作 **8** 张表，从而猜出任何人的年龄（因为一般情况下没有人的年龄超过 **120** 吧）。

大家会问对于一般情形：**N** 个样本中有不止一个坏样本，又该如何设计非序贯的组合检测方案呢？先看一个比较小的例子：**12** 个样本中有 **2** 个坏样本。如果用序贯的组合检测方法，那么用 **7** 次检测就可以确定哪 **2** 个样本是坏的。下图中给出了一个非序贯的组合检测方案，它用 **8** 次检测就可以确定哪 **2** 个样本是坏的（建议大家花点时间，仿照前面 **4** 个样本中有 **1** 个坏样本的例子，验证一下）。

		样　本											
		S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}
检测组	T_1	✓				✓			✓	✓			
	T_2	✓					✓	✓			✓		
	T_3		✓			✓		✓				✓	
	T_4		✓				✓		✓				✓
	T_5			✓		✓					✓		✓
	T_6			✓			✓			✓		✓	
	T_7				✓			✓		✓			✓
	T_8				✓				✓		✓	✓	

一般来讲，一个非序贯的组合检测方案可以用 (**T×N**)-矩阵表示，其中 **T** 表示检测组数，**N** 表示样本数，矩阵的单元(***i***，***j***)为 **1**，表示第 ***i*** 个检测组包含第 ***j*** 个样本；否则为 **0**，表示不包含。大家自然要问了：这个(**T×N**)-矩阵应该满足什么样的性质，**T** 至少要多大（当然不会超过 **N**），其对应的非序贯组合检测方案就可以确定 **N** 个样本中哪 ***d*** 个样本是坏的呢？要回答这个问题需要用到叠加码（superimposed code）的设计方法和理论。这个方法最初是由 W. H. Kautz 和 R. R. Singleton [1]于 1964 年研究如何检索文件时提出来的，后来被 K. A. Bush 等人[2]于 1984 年应用到（非序贯）组合检测方法的设计中。大家对这个方向如有兴趣，还可参阅文献[3,4,5]。

至此，我给大家介绍了与血液检测（直接）相关的五个基本模型及其相应的组合检测方法。下一次，也是最后一次，我将给大家介绍一个血液检测的变形，**伪硬币问题**。

参考文献

[1] W. H. Kautz and R. R. Singleton, Nonrandom binary superimposed codes, **IEEE Trans. Information Theory**, 10 (1964), 363-377.

[2] K. A. Bush, W. T. Federer, H. Pesotan and D. Raghavarao, New combinatorial designs and their applications to group testing, **J. of Statist. Plan. & Infer.**, 10 (1984), 335-343.

[3] F. K. Hwang and V. T. Sos, Non-adaptive hypergeometric group testing, **Studia Scient. Math. Hungarica**, 22 (1987), 257-263.

[4] F. Vakil and M. Parnes, On the structure of a class of sets useful in non-adaptive group-testing, **J. Statist. Plan & Infer.**, 39 (1994), 57-69.

[5] A. G. Dyachkov and V. V. Rykov, a survey of superimposed code theory, **Problems of Control and Information Theory**, 12 (1983), 1-13.