数学无处不在 – 血液检测的数学模型和理论（四）

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）。

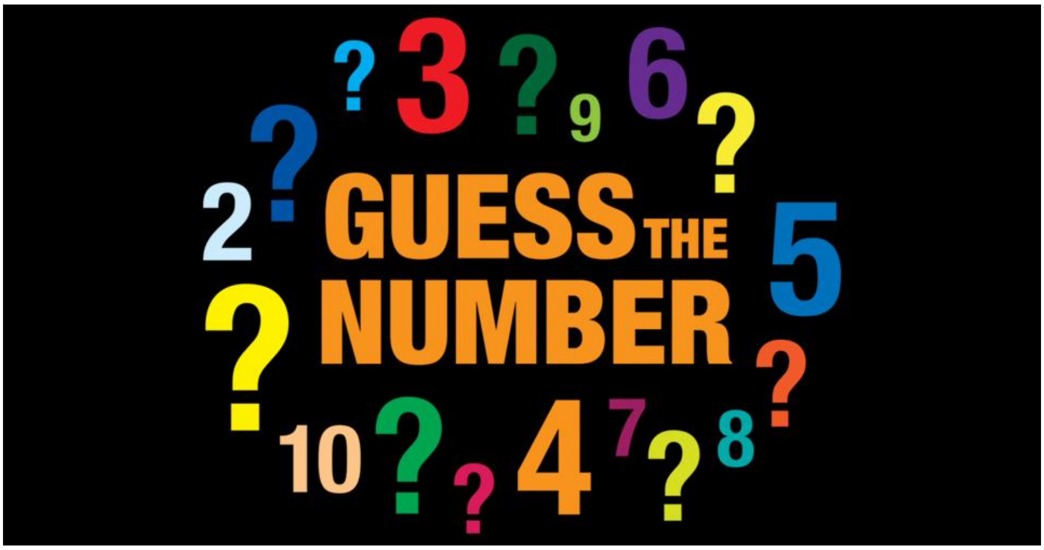
前三次我先后讲了三个模型，**概率模型**：假设事先已经知道所需检测的 **N** 个样本中有 pN 个是阳性的，其中 $0 \le p < 1$；**组合模型**：假设事先已经知道所需检测的 **N** 个样本中有 d 个是阳性的，其中 $0 < d < N$；**竞争模型**：事先既不知道 p 也不知道 d。同时介绍了相应的三个理论结果，在什么情况下，组合检测法需要的检测次数比逐一检测法需要的检测次数少。

今天我给大家介绍**容错模型**：在检测过程中（由于操作或者仪器等因素）会出现错误：对一个没有坏样本的样本组做检测，结果显示阳性（有坏样本），或者对一个有坏样本的样本组做检测，结果却显示阴性（没有坏样本）。

我们是如何能判断出所做过的检测（结果）是有错误呢？一方面，当我们根据完成的检测结果推理出，某个样本（组）应该是阳性的，同时它又应该是阴性的，从而导致了矛盾的结论，我们就判断出已完成的一些检测中必然出现了错误，只不过还无法确定哪些检测出现了差错，哪些没有（比如，对同一个样本做过两次检测，一次显示阳性，另一次显示阴性）。另一方面，根据已完成的检测结果，我们推断出了每一个样本是阳性的或者是阴性的，也没有发现矛盾，但是我们并不能肯定已完成的检测（结果）都是正确的（比如，当对一个样本仅做了一次检测，结果显示阳性；此时无法判断检测结果是否正确）。

在容错模型下，我们又该如何快速并准确无误地检测出所有的坏样本呢？大家很容易想到的重复检测法：按照已有的检测方法做检测，针对每一个待检测样本（组）重复多次检测，如果结果不一致，那么说明检测出现了错误；可以想象，假如每次检测出现错误的概率很小，或者做 k 次检测出现错误的次数不超过 $k/2$，重复检测法是可以准确无误地检测出所有的坏样本。

重复检测法的效率又如何呢？是否会做很多次不必要的检测呢？我先给大家讲一个与此相关的趣题。1976 年著名数学家 S. M. Ulam（S. M. 乌拉姆，1909 - 1984）[1]在其自传中提出了一个猜数问题：

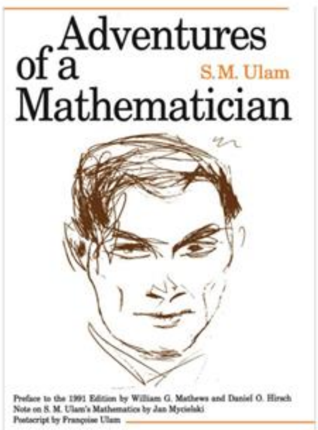


首先，回答者在 1 至 1,000,000（小于 2^{20}）中选定一个自然数为 x。然后，提问者通过向其不断提问，并从得到的回答中猜出 x 的值。比如，提问者可以采用二分策略，首先问“x 是不是在 1 至 500,000 之间？”若他得到的回答为“是”，则他可继续问：“x 是不是在 1 至 250,000 之间？”；若他得到的回答为“否”，则他可继续问：“x 是不是在 250,001 至 375,000 之间？”，等等。显然，提问者最多问 20 个问题，就能猜出 x 的值。当然，提问者也可以问“x 是不是 1？”，“x 是不是 2？”，等等，这样一个数一个数地猜，他至少需要问 999,999 个问题才能确保猜出 x 的值。现在，如果允许回答者说一次或者二次谎，那么提问者至少需要问多少个问题才能确定 x 的值呢？

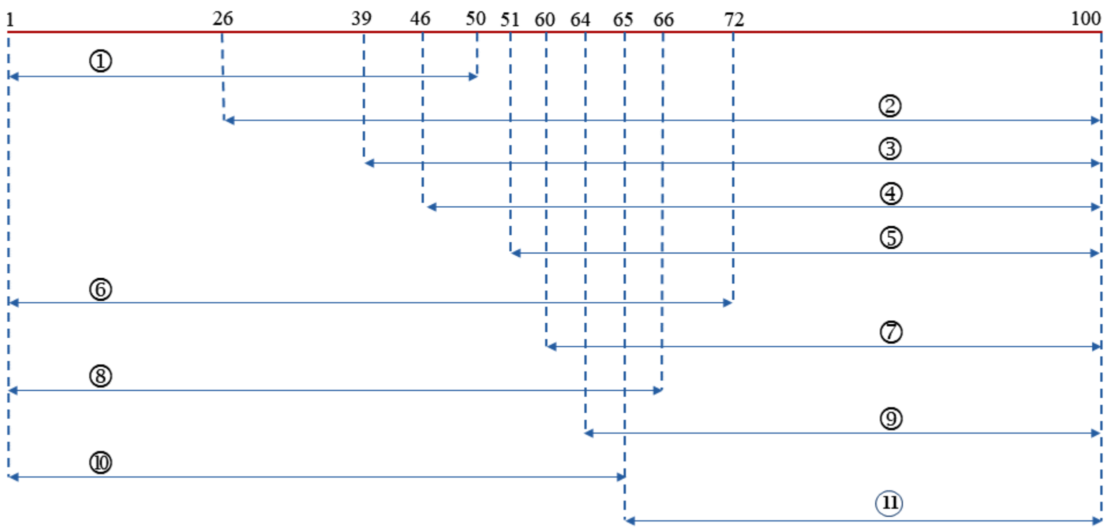
为了讨论简单起见，我们仅允许回答者说谎最多一次。此时，可以采用**重复提问法**：每一个问题问两次，如果两次回答是一致的（表示没有说谎），那么继续提下面的问题；如果两次回答不一致（表示已经说过谎，提问者以后就不能再说谎了），那么同样的问题再问第三次，就得到了正确的回答，然后继续提下面的问题。如此下去，提问者最多问 **41=2×20+1** 次问题就可以猜出 x 的值。提问者能不能用更少的问题还能猜出 x 的值呢？

1984 年 **J. Spencer** [2]对这个问题进行了研究。在其论证中，他允许回答者采用**魔鬼策略**：回答者在提问者开始提问前并不一定需要选定某一个数为 x，实际他只需根据提问者的问题给出回答，使得至少存在一个数，若 x 最初选定这个数时，所有回答除最多一个回答以外都是保持一致的（也就是他最多可能说谎一次）。（实际上，回答者采用魔鬼策略是违反了猜数问题的规则，但是提问者无法判断回答者是否采用了这个策略。）

J. Spencer [2]考虑了一个小例子：**N=100**（小于 2^7）。下面的两个图给出了在提问者与回答者之间进行的 **11** 轮问答过程。注意，当回答者对提问者的第⑤个问题给出了“**Yes**”的回答以后，提问者马上就会发现第⑤个问题和第①个问题的回答中有一个是谎话，只是无法确定哪一个是谎话。不过，提问者能够确认回答者所选定的 x 小于等于 **100** 且大于等于 **46**。因而，他只要用二分搜索法就可以再用 **6** 个问题最终确定 x=**65**。



①	Is $x > 50$?	No.
②	Is $x > 25$?	Yes.
③	Is $x > 38$?	Yes.
④	Is $x > 45$?	Yes.
⑤	Is $x > 50$?	Yes.
⑥	Is $x > 72$?	No.
⑦	Is $x > 59$?	Yes.
⑧	Is $x > 66$?	No.
⑨	Is $x > 63$?	Yes.
⑩	Is $x > 65$?	No.
⑪	Is $x > 64$?	Yes.



现在介绍一下 **J. Spencer** [2]是如何研究：提 k 个问题是否可以确保猜出 x 的值（即使回答者可以说谎一次）？。他引进了一个有序数组(x, L)，其中 x 表示

回答者选择的数，其中 $1 \le x \le N$，L 表示回答者在第几个问题说谎，其中 $0 \le L \le k$（$L=0$ 表示未说谎）。每当回答者针对提问者的一个问题给出“**Yes**”或者“**No**”回答以后，便将(x, L)的可能情形分为两个不相交的集合，**Yes**-集和 **No**-集，分别含有一些可能情形。如果 **Yes** -集包含的可能情形多于 **No**-集，那么回答者就会(贪婪地）选择回答“**Yes**”，否则回答“**No**”。

在上面的例子中，**N=100**，$k=11$。提问者先后问了第①个和第②个问题，回答者分别给出“**No**”和“**Yes**”的回答。当提问者问第③个问题以后，回答者会做出如下分析：

$$
\begin{array}{llr}
(x, L) & 0 < x \leqslant 25, L=2 & 25 \text{ possibilities} \\
 & 25 < x \leqslant 50, L \neq 1,2 & 250 \text{ possibilities} \\
 & 50 < x \leqslant 100, L=1 & 50 \text{ possibilities}
\end{array}
$$

NO CLASS		YES CLASS	
$0 < x \leqslant 25, L=2$	25	$25 < x \leqslant 38, L=3$	13
$25 < x \leqslant 38, L \neq 1,2,3$	$13 \times 9 = 117$	$38 < x \leqslant 50, L \neq 1,2,3$	$12 \times 9 = 108$
$38 < x \leqslant 50, L=3$	12	$50 < x \leqslant 100, L=1$	50
	154		171

在总共 **325** 种可能情形中，**Yes** -集有 **171** 种可能情形，**No**-集有 **154** 种可能情形。回答者会（贪婪地）选择回答“**Yes**”。实际上，提问者的策略就是要选择这样一个问题，其产生的 **Yes**-集和 **No**-集各自含有的可能情形尽可能一样多。可以验证：若提问者第③个问题改为问：“**Is** $x>39$**?**”，则相应产生的 **Yes**-集和 **No**-集所含有的可能情形分别为 **163** 和 **162**。因而，问“**Is** $x>39$**?**”比问“**Is** $x>38$**?**”要更好。

J. Spencer [2]利用魔鬼策略和权函数法，证明提问者只要问不超过 **26** 个问题就可以猜出 x 的值，且 **24** 个问题是不够的，但他留下了一个疑问，用 **25** 个问题可以吗？1987 年 **A. Pelc** [3]通过更加精细的研究，最终证明了 **25** 个问题是可以的，从而彻底解决了乌拉姆的猜数问题。

至此，大家不难看出，乌拉姆的**猜数问题**实际上可以视为容错模型下的**组合检测问题**的一个非常特殊的情形：样本个数是 **1,000,000**，回答者事先选定一个值 i 为 x，也就是第 i 个样本是惟一的坏样本；提问者的针对一组数提一个问题相当于对相应的样本组做一次检测，回答者的回答“是”或“否”表示相应的检测结果是“阳性”或“阴性”(样本组中“含有”或“不含有”坏样本)；回答者说谎相当于检测错误。

最后小结一下我这四次讲的模型和结果：如何用尽可能少的组合检测，在大量的血液样本中准确地检测出所有的坏样本。我介绍的方法基本都属于**序贯方法**：每次检测哪些样本组，要依赖以前的检测及其结果。一个序贯方法如果需要做 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，整个检测过程就可能需要 $n \times t$ 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。下一次我就给大家介绍这种非序贯方法。

参考文献

[1] S. M. Ulam, Adventures of a Mathematician, Scribner’s New York, 1976.

[2] J. Spencer, Guess a number – with lying, Mathematics Magazine, 57 (2) (1984), 105-108.

[3] A. Pelc, Solution of Ulam’s problem on searching with a lie, Journal of Combinatorial Theory, Series A, 44 (1987), 129-140.