对数秩“Logrank”检验名字的由来

Mai Zhou*

在生物统计，医学统计，生存分析中，Logrank 检验是最常用的一个统计检验。一般它用来比较两个或多个母体是否相同（新药是否比旧药好）。它可以应用于右删失数据，而且不需要假定母体为某个参数分布族（比如 Weibull 分布，等等）。中文译为对数秩检验。百度搜索“对数秩检验”，有超过 50 万条网页之多。用 Google 搜索“log rank test”，有超过 7 亿条。

我们在这儿要澄清关于 Logrank 这个名字由来的一些（在中文网页上流传的）错误传说。以及用类比电子游戏比赛来理解这个检验方法。也提供了 Logrank 这个名字的原因。

1 Logrank 检验最早由 Peto and Peto 两兄弟在 1972 年命名

有兴趣的读者可以查看他们的原文 [1]。在他们之前，之后这个检验方法也被其他人提出来过：所以这个检验也被称为（可能还有更多）

- 1. Mantel–Haenszel 检验，（作为一系列 2×2 表之和）。
- 2. Cox 比例风险回归中的“得分检验”，（偏似然的导数）。
- 3. 在数据没有删失的情况下，Savage 检验，（对于 Lehmann 对立假设的最佳线性秩检验）。
- 4. 两个 Nelson-Aalen 估计的累积加权差别。（和点过程鞅联系最密切）。

在 Peto and Peto 文章之后，“Logrank”这个名字就被大家叫开了，普遍接受了。它的意思就是“数据的秩的对数”。这里 Log 就是指对数函数。不像有些网页上声称的那样“Log”不是指对数函数，而是指日志记事本（logbook）的意思。计算机登录/注销，英文是“Login/logout”这和 Logbook 意思相似。

*Dr. Bing Zhang Department of Statistics, University of Kentucky, Lexington, KY 40536 Email: maizhou@gmail.com

那为什么 Peto 两兄弟要把它叫成 Logrank 呢? 他们在 1972 年的文章 [1] 里给出了他们的理由。我们在这里给出一个简单的例子说明。这个解释与（概率统计中的）鞅的概念密切相关，而且用这个理解很容易解释在具体使用 Logrank 检验中，病人可以迟到，早退，而不影响检验的公平有效性。

2 Logrank 检验的电子游戏理解

假设有两支队伍比赛打电子游戏。不妨记为 A 队和 B 队。A 队有 n 人，B 队有 m 人。游戏规则如下：

- 1. 若要参赛，必须拿出 1 块钱放到桌面上（可以理解为买票）。
- 2. 游戏的目的是坚持不被怪物吃掉，坚持（生存）得越久，可以赢取越多的钱。
- 3. 先假设所有人同时参加游戏，当某一人被怪物吃掉时，他/她的 1 块门票钱要平分给所有在当时仍然没被吃掉的人（再加上她/他自己），都有一份，然后离场。

  比如，第一个被吃掉的人，只能拿回 $1/(n+m)$ 块钱，却付出了 1 块钱。总收益 $= 1/(n+m) - 1$。

  第二个被吃掉的人，拿回了 $1/(n+m) + 1/(n+m-1)$, 也付出了 1 块钱。总收益 $= 1/(n+m) + 1/(n+m-1) - 1$。

  如此等等。

  坚持到最后第二个（亚军），总收益 $= 1/(n+m) + \cdots + 1/2 - 1$。

  坚持到最后的人（冠军），他/她的门票钱，只与他/她自己分，也就是最后被吃掉时，拿回 1 块钱。也付出了 1 块钱。总共有收益 $1/(n+m) + 1/(n+m-1) + \cdots 1/2 + 1 - 1$。

  其实不必要求所有人同时参加游戏，这时上面第三条可以改为

- 3'. 不论有多少人正在游戏，当某一人被怪物吃掉时，他/她的 1 块门票钱要平分，给所有在当时正在游戏，仍然没被吃掉的人（再加上她/他自己），都有一份，然后离场。

很明显，如果 A 队队员们技术普遍比 B 队好，平均生存时间长，那么 A 队的总收益将大概率为正数。反之亦然。

定义：比较 A 队与 B 队的净收益就是 Logrank 检验。由于这是一个 “零和游戏”，所以只看一个队的净收益就可以了。当然，在做统计检验的时候，还要将此数除以它的方差。而方差估计有几种方法，结果大同小异。我们就不深入讲解了。

定理：这是一个公平游戏。

证明：什么是 “公平游戏”？就是在每个人技术水平一样的前提下，每人的付出（1 块钱门票钱）和未来的期望收益（平均收益）相等，也是 1 块钱。

比如说，只有一个人在游戏，付出 1 块，也必定拿回 1 块。

如果有两个人在游戏，付出 1 块，未来收益期望是：0.5 概率可能收回 0.5 块，0.5 概率可能收回 1.5 块；所以收益期望也是 1 块。（有 50% 可能输/赢）。

如果有三个人，因为大家技术水平一样，所以此人得第三/第二/第一的概率都是 1/3. 收益分别是 1/3, 1/3+1/2, 1/3+1/2+1。算一下期望，也是 1 块。

其余情况类推。或者用归纳法。

早退。任何人在任何时候（当然，假定还没被怪物吃掉）要走人，可以拿回 1 块钱就走人。对应在临床试验时，有些志愿者因为各种（与试验无关的）原因不得不停止试验。为什么可以拿回 1 块？因为如果此人继续参加游戏，其未来收益期望是 1 块。

迟到。任何人在任何时候要参加游戏，只需拿 1 块钱放桌上，就可以立即加入。这就像在医学临床试验时，需要招募志愿者，是陆陆续续加入的。

这个游戏可以在任何时候完全停止，就像大家一起早退。或者就像临床试验被叫停，或者到了预先计划的 5 年期限，等等。

定理：即使有人迟到，早退，可以被叫停，这还是一个公平游戏。

证明：假设在某个时刻有 k 个人在游戏，由于大家技术水平一样，此人有 $1/k$ 的概率是这些人中最先出局的，这时收益为 $1/k$。如果不是最先出局，则有收益 $(1/k+1)$ [1]。这个理由对于每个人都适用。未来收入期望：

$$\frac{1}{k}\times\frac{1}{k}+\left(\frac{1}{k}+1\right)\times\left(1-\frac{1}{k}\right)=1 \ 。$$

也就是说，这个游戏是每时每刻都公平的，不是非得进行到底才公平。每时每刻的未来收益期望总是 1 块（所以门票也总是 1 块）。每时每刻的未来条件期望总是零，这就是一个鞅，参见 [2].

3 Logrank 名字

现在回到 “Logrank” 这个名字的由来。首先数学里有如下的约等式：

$$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{n}\approx\log(n) \ 。$$

我们由上面的游戏规则知道，假如有 $(n+m)=N$ 个人参加游戏，那第 r 个被淘汰的人具有收益：

$$\frac{1}{N}+\cdots+\frac{1}{N-r+1}-1\approx\log(N)-\log(N-r)-1=-\log\left(1-\frac{r}{N}\right)-1$$

[1]此人可以在拿了 $1/k$ 之后立即退赛（拿回门票钱）。或者休息一下再加入游戏。

如果此人不是被淘汰，而是由于别的原因自己退赛（删失数据），那此人的收益是上面的数, 不过没了后面的 -1。(退赛可以拿回门票钱)。

如果反过来算，即第 1 名（冠军）是最后被淘汰者，等等。那么第 r 名有收益

$$\frac{1}{N}+\frac{1}{N-1}+\cdots+\frac{1}{r}-1\approx\log(N)-\log(r-1)-1=-\log\left(\frac{r-1}{N}\right)-1$$

如果此人在还剩 r 人时主动退赛，则收益为上式去掉 -1。

显然，上面的约等式在 $\log(\cdot)$ 函数的自变量比较小的时候，是不对的/没有定义的。但是它形象的说明一个参赛者的得分/收益，由某个秩的函数决定，而这个函数(近似的）与 Log 函数有关。

检验统计量就是 A 队的总得分/收益。

4 两种计算队伍总收益的方法

前面说了，Logrank 检验就是看一个队的总收益。计算一个队伍的总收益可以有两种方法。很显然两方法殊途同归。

1 （按队员计算）每个参赛人员自己记录自己的收益。完赛后 A 队长召集所有队员将每个队员的收益相加。

2 （按时间计算）在每一次有人被怪物吃掉时，计算 A 队全队在此次事件中的收益，然后将每次事件的收益相加。

前面提到的 Mantel–Haenszel 检验，就是在每当有人被怪物吃掉时，做一个 2×2 表来计算 A 队全队在此次事件中的收益与 B 队收益之差。常用的软件 SAS 也采用了这个计算方法 [3]。这个方法从一个队伍的角度出发计算收益。

我们在前面描述的电子游戏方式，是用第一种方法，从每个个人的角度出发来计算全队收益。

References

[1] Peto, R. and Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society Series A* Vol. 135 No. 2 pp. 185-207. http://www.jstor.org/stable/2344317

[2] Lan, K.K.G., and Lachin, J.M. (1995). Martingales without tears. *Lifetime Data Anal*, Vol. 1, 361–375. https://doi.org/10.1007/BF00985450

[3] Allison, P. (2010). Survival Analysis Using SAS: A Practical Guide, 2nd Ed. SAS Institute.