Stein estimator: 逝去的悖论 *

传说中，有一个叫 Stein parodox 的，不知您听说过没有。直接在维基上查“Stein”，条目很多，叫”Stein” 的人名、河流、山脉、村庄，应有尽有，甚至月亮上有一座环形山，也命名为“Stein”。这里所谈的“Stein”，是一位美国统计学家 –Charles Stein。他的相关维基词条还有：James-Stein Estimator、Steins lemma、Stein'smethod、Stein paradox、Stein's unbiased risk estimate 和 Stein's loss。一个人在维基上能有这么多词条，足以说明这个人的非凡之处。今天，就用通俗的语言，来介绍一下当年被视为悖论的“James-Stein Estimator”。(James, 全名是 Willard James，是 Stein 的一名非全职弟子，与老师合写的（James & Stein1961）是他发表的唯一一篇统计文章。为了方便起见，我们把 James-Stein 估计简称为 Stein 估计。）

Figure 1: Charls Stein

Stein's Estimator 缘起于 Stein(1956), 在 James & Stein(1961) 中得到完善，自其提出后，广受质疑，被人们当做不起眼的悖论。是 Efron, 没错，就是 Efron Brandly(Bootstrap、LAR 算法等等的发明人)，为其找到了合理的解释 –Emprical Bayes。Efron 的《Large-Scale InferenceEmpirical Bayes Methods for Estimation, Testing, and Prediction》第一章：“Empirical Bayes and the James-Stein Estimator” 详细解释了 Stein estimator 的贝叶斯含义，其解释给人以拨云见日、别有洞天的感觉，让我们感悟到统计的魅力。我们就从 Efron 的著作出发，说明什么是 Stein's Estimator，为什么它被定义悖论，又该如何解释这个悖论。

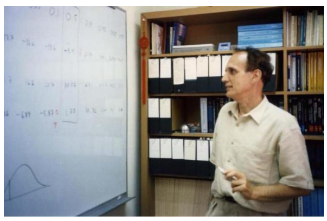

Figure 2: Efron

*本文作者高磊，天津财经大学统计系 2013 级博士生.

这一赛季的棒球联赛已经进行一段时间[1]，我们搜集到 18 个球员的技术数据，整理如下表。稍微解释一下，共统计了 18 名球员，这些球员都打了 45 棒（AB)，统计了他们击中的次数（hits)。例如，Clemente 共打了 45 棒，击中了 18 次；F Robinson 也打了 45 棒，击中了 17 次；Alvis 也打了 45 次，只击中了 7 次。

Name	hits/AB	$\hat{\mu}_i^{MLE}$	μ_i	$\hat{\mu}_i^{JS}$
Clemente	18/45	0.400	0.346	0.294
F Robinson	17/45	0.378	0.298	0.298
F Howard	16/45	0.356	0.276	0.285
Johnstone	15/45	0.333	0.222	0.280
Berry	14/45	0.311	0.273	0.275
Spencer	14/45	0.311	0.270	0.275
Kessinger	13/45	0.298	0.263	0.270
L Alvarado	12/45	0.267	0.210	0.266
Santo	11/45	0.244	0.269	0.261
Swoboda	11/45	0.244	0.230	0.261
Unser	10/45	0.222	0.264	0.256
Williams	10/45	0.222	0.256	0.256
Scott	10/45	0.222	0.303	0.256
Petrocelli	10/45	0.222	0.264	0.256
E Rodriguez	10/45	0.222	0.226	0.256
Campaneris	9/45	0.200	0.286	0.252
Munson	8/45	0.178	0.316	0.247
Alvis	7/45	0.156	0.200	0.242
Grand Average		0.265	0.265	0.265

Table 1: Batting

摆在面前的问题是，能不能估计各位球员的击中率，以此作为该球员在未来比赛中击球率的预测值。从统计的鼻祖伯努利开始，我们就熟悉，可以用一个简单的除法比率作为击中率的估计值。以 Clemente 为例，他打了 45 次，击中 18 次，所以他的击中率是 $18/45 = 0.400$, 其他球员都类似处理。戴上 Fisher 的眼镜来看，我们做的是极大似然估计，极大似然估计被 Fisher 标榜为估计的“绝对准则”（Absolute criterion）[2]。Fisher 是一名富有战斗力的学者（一战时准备入伍，但因视力不合格而被拒绝)，他的“绝对准则”不仅向 Karl Pearson 的“矩估计”开了一炮，还把一大批后生给吓唬住了，以为极大似然估计是统计估计的“圣杯”。

长江后浪推前浪，世上本没有什么绝对准则，包括统计估计方法。Stein ，二战中在为军方服务的统计实验室工作，被授予“上尉军衔”, 在二战结束后的 1950 年提出了 Stein 估计的思想。Stein 认为，当我们对一个样本的期望进行估计时，均值估计，也可以说是极大似然估计，是个好的选择（注意到，上边我们对击中率估计本质上就是求均值)。返回到例子中，如果只有 Clemente 的数据，也就是说，只有单个样本（这个样本容量为 45，其中 18 个 1,27 个 0)，没有其他运动员数据，这时候用 $18/45 = 0.400$ 做估计是木有问题的。Stein 还证明，当样本个数增为两个，比如我们又有了

[1]棒球的例子，来源于 Efron,B.;Morris, C.(1977).”Stein’s paradox in statistics”

[2]“绝对准则”出现在 Fisher（1912）”On an Absolute Criterion for Fitting Frequency Curves” 标题中。

Figure 3: Fisher

F Robinson 的数据，用极大似然方法分别来估计也是靠谱的，估计值分别为 0.400 和 0.378。Stein 又证明**当样本个数超过 2 个时，不管你信不信，极大似然估计就不是最好的选择！**空口无凭，Stein 给出了 Stein estimator。令人“大跌”Fisher 眼镜的是，这个统计量的表现确确实实优于极大似然估计。这里就不再啰嗦 Stein 估计公式了[3]，但我们仍然可以一睹 Stein 估计的“芳容”：

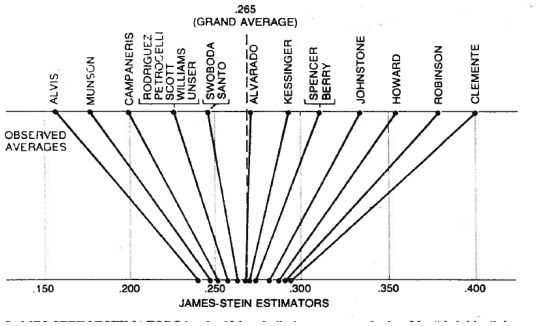

Figure 4: Shrinkage

一图胜万言。上侧部分，是各个运动员击中率的极大似然估计，对应于 Table1 中 $\hat{u}_i^{MLE}$ 列；图中间有一条竖直虚线，指的是所有运动员击中率的平均估计，0.265，在 Stein estimator 中，这是非常重要的一个值；所谓的 Stein 估计，就是在极大地然估计的基础上，向整体水平（0.265）进行收缩（shrinkage), 收缩后的估计值（对应于 Table1 中 $\hat{u}_i^{JS}$ 列就是下侧直线的点）。

下侧的值，真的比上侧的值，对未来估计更准确吗？答案是真的！在 Table1 中，还给出了所有运动员，未来打球真实的击中率 μ_i。将两种估计与真实值的差距取平方，再求均值（取名为均方误差），我们发现，Stein 估计的均方误差比极大似然估计要小许多，说明其估计值与真实值偏离要小，用 Stein 估计的结果与真实值的差异竟然只有极大似然估计的 30%，用 Efron 的话说，“The gains from using $\hat{u}^{JS}$ can be substantial”。

$$\sum_{1}^{18}\left(\hat{u}_i^{JS}-\mu_i\right)^2 \Big/ \sum_{1}^{18}\left(\hat{u}_i^{MLE}-\mu_i\right)^2 = 0.28$$

在那个以 Fisher 为代表的频率学派如日中天的时代，Stein 估计被看做一个悖论，不过仅仅被看做一个悖论。人们很难想象，在对一个样本（一个运动员）进行估计的时候，干嘛要把另一个样本（其他样本）拉扯进来，并对该样本估计造成影响。依 Stein 的逻辑，在对棒球球员击中率进行估计的时候，如果你恰好还有 NBA 联赛三分球投球命中数据的话，那么你最好也把这些数据加进

[3]有兴趣的读者可查阅 《那些年，我们追的 EB》

来，势必会改善估计的结果。在统计人看来，这不免有些荒谬，为什么棒球运动员的击中率要和篮球运动员的三分命中率联合起来？这给人一种很不舒服的感觉，这也是 Stein 估计被当做悖论的缘由所在。我们看看陈希孺老先生在《数理统计简史》中是如何描绘这件事的,

Figure 5: 院士陈希孺

.....

本来估计 $a_1,\cdots,a_p$ 是 p 个不相干的问题。照常理，估计 a_i 只应到到与之相关的样本 $x_{i1},\cdots,x_{in}$，而在这个场合下 $\bar{x}$，已经是一个良好的估计。现在 *Stein* 的结果告诉我们说，情况并非如此，**在估计 a_i 时，除了使用 $x_{i1},\cdots,x_{in}$ 外，还要使用另外 $p-1$ 组与之不相干的样本，才能得到更好的结果。这个说法与常理相违背**.

这个结果的深刻含义在于：它显示了数学理论与使用考虑之间的一种不合拍，因而使人对这种理论的有效性提出了疑问。毕竟**统计学是一门实用学科**，一个问题，从模型提法、优良性准则到数学论证，不论看上去多么合理，最后还得落实到应用上的合理性这一条。对 *Stein* 这个结果从实用层面来看，不会动摇人们对习以为常的估计 $(\bar{X}_1,\cdots,\bar{X}_p$ 的信赖，而是反过来，**对平方误差损失，对用风险函数衡量一个估计的优良性这些基本出发点的合理性提出质疑。**

.....

用平方误差损失以及风险函数来评价估计的优良性是 Abraham Wald《统计决策论》的核心内容，可以看做是对 Fisher UMVU 估计（无偏限制下、方差一致最小估计）的重大突破。Stein,Wald 的得意门生，正是在老师的分析框架证明，当样本量超过 3 个是，极大似然估计并不是最优的。依《简史》的观点看来，不仅 Stein 的结论是数学游戏，就连 Wald 的统计决策论也不靠谱。

不过，《简史》的分析逻辑的出发点“实用至上”，这么说有点不太舒服，套用一句官话“**实践是检验真理的唯一标准**”。我非常认同这种看法，统计学到头来还是要服务现实社会，统计学家不能自娱自乐。然而，即使从“实践”的观点出发，Stein 估计真的是一个纯粹的理论幌子吗？在 Stein 估计提出后，有些人给予了较高的评价，Neyman 就是其中之一，把它当做统计理论的重大突破；也有一些人把它当做一个悖论，认为它毫无现实意义。江山代有才人出，1970 年，Efron 在《科学美国人》上发表《Stein Paradox in statistics》，文章开头首先指出，Stein paradox 该是寿终正寝的时候了，文中几个漂亮的例子不禁让人们对 Stein 估计着迷，不过文章的亮点是最后 Efron 对 Stein 现象的（经验）贝叶斯解释，顿时给人一种“柳暗花明又一村”的感觉。

Efron 的（经验）贝叶斯解释，这里不再赘述，兴趣读者可见[4]。其中的关键假设就是，我们关心的样本之间往往是具有某种联系的，这是应用贝叶斯假设的一个前提。这里的“某种联系”，说

[4]《Large scale inference》第一章”Emprical Bayes and James-Stien Estimator”

白了，就是我们认为考察的 18 名运动员的击球率比较接近他们的共同平均水平（必须承认这一点）。也就是说，当我们对一个样本进行估计时，我们需要首先考察其他样本的信息，这就是所谓的”Leaning from the Experience of Others”, 如下图所示：

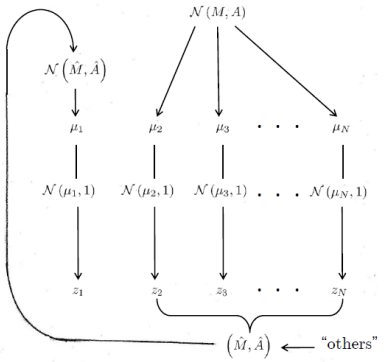

Figure 6: Learning Experiences from others.

Efron 认为，在科学大生产时代，相近水平的类似样本估计很普遍。比如要同时估计上万个基因水平，这些样本之间往往具有某种内在联系，因此可以利用其它样本对当前样本的估计修正。在棒球的例子中，不考察其他运动员的信息又怎么样？举个极端的例子[5]：在这十八名球员之外，再引入一位“双球”运动员：Frank O’Connor，他一年赛季当中，只打了两个球，巧的是，他两次都打中了。那么该怎么估计他的击中率？如果用 $2/2 = 100\%$ 作为他击中率的估计值，用 Efron 的话，“This is a vey silly estimate”。O’Connor 在这个赛季只打了两个球，这个数据量太少，必须通过其他球员的数据对 O’Connor 的估计校准，单用 O’Connor 的数据做估计，误差太大，就算这个数据是真实的，它顶多说明这两次他发挥不错。更精确的估计是在其他球员平均水平（0.265）基础上，稍微向高的方向偏一点，比如 0.272 ，这也就相当于从极大似然估计向平均水平进行收缩。

极大似然估计的收缩有时是剧烈的，如下图所示，极大似然估计排第一的一个样本估计，通过 Stein 估计，竟然收缩到了 12 名，其原因就是该样本的样本量较少，估计误差偏大，极大似然估计很不可信，必须借助于其他样本信息对这个样本的估计进行修正。这个道理其实可以算是一个常识，而极大似然估计往往会罔顾事实。

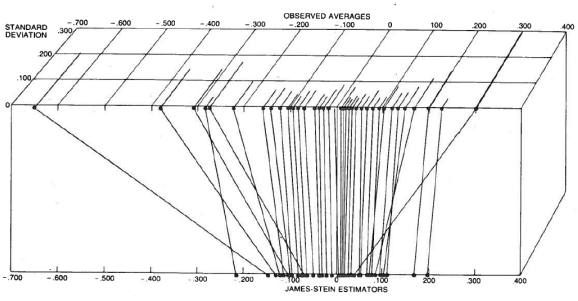

Figure 7: Shringkage

从极大似然估计向平均水平收缩，有着非常重要的现实意义。延续上面的例子，如果要对 19 名棒球运动员排名，并按排名前后发放奖金，第一名 190,000，第二名 180,000，依次递减。那

[5]Efron,B.;Morris, C.(1977).”Stein’s paradox in statistics”

么 O'Connor 该排第几？如果不发钱，对其他运动员而言，他排第几真的无所谓。但是当估计涉及个人利益，每个人都会刷刷地打起小算盘。O'Connor 拿 190,000（按极大似然估计，击中率为 $2/2 = 100\%$），那 18 位定会满腔怨言；而按 Stein 估计，O'Connor 的名次大概排中游，更容易被人接受，但就算这样估计，也会受到别人的冷笑，认为他是个机会主义者，钻了空子，仅凭两个球就拿到了 100,000。

惯常理解，硝烟战场上泣血搏杀的战士被当做英雄。可是在一门学科的版图上自由驰骋、开疆扩域的智者是不是也算是英雄呢？20 世纪，是一个统计英雄辈出的时代。Karl Pearson,Fisher,Neymann,Stein,Efron 是否堪当此喻，后人自有评说。也许现在还不是华山论剑自我标榜的时刻，我们所能做的，就是爬到巨人的肩膀，看向更远的远方！

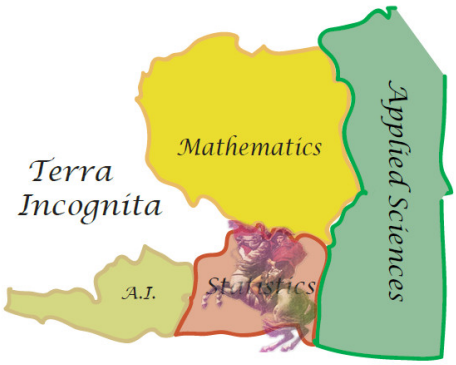

Figure 8: The greater world of statistics.

参考文献：

- 陈希孺.《数理统计学简史》.2002. 湖南教育出版社.
- Stein, C. (1956), "Inadmissibility of the usual estimator for the mean of a multivariate distribution", Proc. Third Berkeley Symp. Math. Statist. Prob. 1, pp. 197–206, MR 0084922, Zbl 0073.35602
- James, W.; Stein, C. (1961), "Estimation with quadratic loss", Proc. Fourth Berkeley Symp. Math. Statist. Prob. 1, pp. 361–379, MR 0133191
- Morris H. DeGroot A Conversation with Charles Stein, Statistical Science, Vol. 1, No. 4 (Nov., 1986), pp. 454–462.
- Efron, B.; Morris, C. (1977). "Stein's paradox in statistics". Scientific American 236 (5): 119–127.
- R. A. Fisher (1912). On an Absolute Criterion for Fitting Frequency Curves". Messenger of Mathematics 41: 155–160.
- Efron. (2012) Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.