Silkman Statistical Journal

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

最大似然的史诗故事

Stephen M. Stigler*　　**翻译：林绪虹**†

摘要

从表面上看，最大似然法的思想一定是史前的：早期的狩猎者和采集者可能没有使用“最大似然法”这个词来描述他们选择在哪里以及如何狩猎和采集，但如果他们的方法被这样描述，他们很难相信会感到惊讶。这似乎是一个简单的，甚至是无懈可击的想法：谁会站出来支持最小似然法，甚至是中等似然法？然而，这个主题的数学历史表明，这个“简单的想法”其实一点也不简单。约瑟夫·路易斯·拉格朗日、丹尼尔·伯努利、伦纳德·欧拉、皮埃尔·西蒙·拉普拉斯和卡尔·弗里德里希·高斯只是探索这个主题的部分人，他们的方式并不总是我们今天认可的。在本文中，回顾了从费舍尔之前到吕西安·勒卡姆论文发表时期的历史。在此过程中，介绍了费舍尔 1930 年未发表的对最大似然估计的一致性和效率条件的描述，并讨论了他的三个证明的数学基础。尤其是，Fisher 的信息不等式的推导被认为是从他的方差分析工作中衍生而来的，而他后来通过估计函数的方法则从欧拉的齐次函数关系中衍生而来。本文回顾了人们对 Fisher 工作的反应，并得出了一些教训。

Keywords: R. A. Fisher，Karl Pearson，Jerzy Neyman，Harold Hotelling，Abraham Wald，最大似然，充分性，效率，超效率，统计学史.

1. 简介[1]

19 世纪 60 年代，一小群年轻的英国知识分子成立了他们所谓的 X 俱乐部。这个名字被用作未知的数学符号，计划是每月聚在一起吃一次晚餐，让谈话带他们去碰碰运气的地方。这个团体包括达尔文生物学家托马斯·亨利·赫胥黎（Thomas Henry Huxley）和社会哲学家兼科学家赫伯特·斯宾塞（Herbert Spencer）。1870 年左右的一个晚上，他们在伦敦的 Athenaeum 俱乐部共进晚餐，那天晚上的一次谈话让在场的人大吃一惊，以至于在场的人重复了好几次。弗朗西斯·高尔顿（Francis Galton）没有出席晚宴，但他从三位在场的人那里听到了不同的叙述，并把它记录在自己的回忆录中。正如高尔顿所报道的那样，在谈话的暂停期间，赫伯特·斯宾塞说：“你可能想不到，我曾经写过一部

[1]这是《统计科学数理统计研究所》2007 年第 22 卷第 4 期第 598-620 页发表的原始文章的电子重印本。重印本在页码和排版细节上与原文有所不同。

悲剧。”赫胥黎立即回答说：“我知道这场灾难。”斯宾塞说这是不可能的，因为他之前从未谈论过它。赫胥黎坚持说。斯宾塞问那是什么。赫胥黎回答说：“一个美丽的理论，被一个肮脏、丑陋的小事实扼杀了。”（Galton，1908 年，第 258 页）

乔·霍奇斯的《丑陋的小事实》（1951 年）

$$T_n = \bar{X}_n \qquad \text{if} |\bar{X}_n| \geq \tfrac{1}{n^{1/4}}$$
$$= \alpha \bar{X}_n \quad \text{if} |\bar{X}_n| < \tfrac{1}{n^{1/4}}.$$

那么，如果 $\theta \neq 0$，则 $\sqrt{n}(T_n - \theta)$ 渐近于 $N(0,1)$，如果 $\theta = 0$，则渐近于 $N(0, \alpha^2)$。

如果 $\alpha^2 < 1$，则 T_n 对于 $\theta = 0$ 是“超高效的”。

图 1: Joseph L. Hodges, Jr. 提出的超高效估计示例。该示例在 1951 年的讲座中提出，但首次发表于 Le Cam (1953)。这里 $\bar{X}_n$ 是从 $N(\theta, 1)$ 总体中随机抽取的大小为 n 的样本均值，其中 $n \operatorname{Var}(\bar{X}_n) = 1$ *all* n, *all* θ (Bahadur, 1983; van der Vaart, 1997)。

赫胥黎对科学悲剧的描述特别适合讲述最大似然法的历史。最大似然理论确实非常美妙：一种概念上简单的方法，可以解决大量问题。该理论提供了一个简单的方法，旨在为所有参数问题及其他问题提供最佳解决方案，不仅承诺提供最佳估计，而且还提供简单的通用准确性评估。所有这些都不需要指定先验概率，也不需要复杂的分布推导。此外，它可以在现代计算机中实现自动化并扩展到任意维度。但正如赫胥黎对斯宾塞未发表的悲剧的讽刺一样，有些人认为该理论“被一个令人讨厌的小事实扼杀了”，最著名的是约瑟夫·霍奇斯在 1951 年提出的优雅简单的例子，指出存在“超高效”估计（具有比最大似然估计更小的渐近方差的估计）。见图 1 。然后，就像罗马斗兽场中致命伤的奴隶，或西班牙斗牛场中致命伤的公牛一样，该理论再次被其他人用不一致的最大似然估计的巧妙例子多次击溃。

最大似然理论的整个故事比这个简单的叙述要复杂得多，也少了些悲剧色彩。最大似然理论的历史更像荷马史诗，长期的和平时期被一些小规模的攻击打断，这些攻击最终演变成大规模的战争；胜利与悲剧交织在一起，所有这些都被一些英雄人物所主导，即使他们没有英雄气质。尽管过去动荡不安，最大似然理论经受住了无数次攻击，仍然是一个美丽的理论，尽管它变得越来越复杂。我打算回顾一下这段历史，先概述一下早期的概念问题，然后仔细看看 20 世纪 20 年代和 30 年代的大胆主张，以及为支持这些主张而提出的一些未发表的早期论据。

2. 最大似然法的早期历史

到十八世纪中期，自然哲学家们似乎普遍认为，观测误差问题可以用数学描述。对于这种描述的一些要素，人们基本达成了一致：由于缺乏更好的假设，误差被认为同样可能为正或负，大误差出现的频率预计要低于小误差。事实上，人们普遍认为它们的频率分布遵循平滑的对称曲线。甚至观察者的目标也达成了一致：虽然使用的词语各不相同，但观察者会寻找观察对象最可能的位置，无论是恒星赤纬还是大地坐标位置。但在为数不多的认真尝试处理这个问题的过程中，细节发生了重大变化。事实证明，要得出一个精确的公式来包含这些元素、涵盖有用的应用，并允许进行分析，是相当困难的。

早在 18 世纪 50 年代，托马斯·辛普森和贝叶斯，以及 1760 年约翰·海因里希·兰伯特就对这一问题作出了明智的评论，但第一次对我们的主题进行认真的攻击是 1769 年约瑟夫·路易·拉格朗

日 (Stigler, 1986, Chapter 2; 1999, Chapter 16; Sheynin, 1971; Hald, 1998, 2007)。拉格朗日假设观测值按照多项分布围绕期望平均值变化，他在分析中展示了如果将不同可能值的相对频率用作概率值，则一组观测值的概率最大。用现代术语来说，他发现多项概率的最大似然估计是样本相对频率。他得出结论，期望平均值的最可能值是从这些概率中找到的平均值，即观测值的算术平均值。直到那时，与现代实践相反，拉格朗日才引入了多项概率遵循对称曲线的假设，因此他只剩下当误差概率遵循曲线时找到算术平均值的概率分布的问题。他通过引入和使用“拉普拉斯变换”解决了几个例子。通过在得出概率估计之后引入曲线形式的限制，拉格朗日的分析产生了一个奇怪的结果：尽管从最大似然开始，但总是得出矩法估计！(Lagrange, 1776; Stigler, 1999, Chapter 14; Hald, 1998, page 48.)

大约在同一时间，丹尼尔 • 伯努利（Daniel Bernoulli）以两种截然不同的方式考虑了这个问题。首先，在 1769 年，他尝试使用假设曲线作为权重函数，以便加权，然后迭代地重新加权并平均观测值。这非常类似于一些现代稳健 M 估计。其次，在 1778 年（可能是在他看到了拉普拉斯 1774 年的回忆录，其中包含贝叶斯分析公式之后），伯努利彻底改变了他的观点，并使用同一条曲线作为单个观测值的密度。他将这些密度相乘，并寻求观测量的真实值，即使乘积达到最大值的值 (Bernoulli, 1769, 1778; Stigler, 1999, Chapter 14; Laplace, 1774)。

这些尝试以及当时的其他尝试主要是理论探索，并没有吸引到太多实际应用或进一步发展。尽管它们都使用了可以很容易地翻译成现代英语的短语“最大似然”，在某些情况下甚至可以用最大似然来辩护，但在任何情况下，都没有为它们或它们的性能提供合理的辩护。

最多可以发现的是，表面上调用的是得出的值“最有可能”，因为它使唯一可见的概率（观察到的数据的概率）尽可能大。

在这些早期处理方法中，从哲学角度来看，最有说服力的是高斯在 1809 年发表的第一篇关于最小二乘法的文章（Gauss, 1809）。高斯和 1778 年的丹尼尔 • 伯努利一样，采用了拉普拉斯的解析公式，但与伯努利不同的是，高斯明确援引了拉普拉斯的贝叶斯观点，对未知数使用均匀先验分布。拉普拉斯随后寻找（并找到了）后验中位数（最小化后验预期误差），而高斯选择了后验模式。根据现代最大似然法和正态分布误差，高斯采用了最小二乘法。由于分析简单易行，这种方法在 19 世纪非常流行。到 19 世纪末，这种方法有时被称为高斯方法，并且成为许多教科书的主要内容，通常没有明确援引高斯认为需要证明该方法合理性的均匀先验。

3. 卡尔 · 皮尔逊和 L. N. G. 菲隆

19 世纪，估计理论总体上仍停留在拉普拉斯和高斯留下的水平，尽管经常退回到较低的水平。关于最大似然，在高斯 1809 年发表论文之后，最重要的事件发生在新世纪的前夕，卡尔 • 皮尔逊和路易 • 拿破仑 • 乔治 • 菲隆在 1898 年发表的《伦敦皇家学会会刊》上发表了一篇长篇回忆录 (Pearson and Filon, 1898)。1898 年出版的回忆录在历史上占有一席之地，更多的是因为它最终似乎提出了什么，而不是它取得了什么成就。两位作者考虑了一个非常普遍的估计问题环境——一组多元观测值，其分布取决于可能很大的一组待确定的常数。他们没有将常数称为参数，但现代读者很难从其他角度看待它们，尽管仔细阅读回忆录会发现它缺乏 Fisher 20 多年后引入的参数观点（Stigler，2007 年）。

皮尔逊和菲隆的主要结果（用现代术语来表达）来自取似然比（观测数据的频率分布与对相同数据进行评估的频率分布的比率，但常数略有扰动），在多元泰勒展开式中展开其对数，然后用其期望值

近似系数，并声称所得表达式给出了估计常数时产生的误差的频率分布。他们错误地取了系数的极限，实际上使用了一种完全不依赖于所用估计方法的程序，最多对最大似然估计有效，但他们没有意识到这一事实。他们的最后一步采用了高斯方式的隐式贝叶斯步骤。当忽略三次和更高阶项时，他们的公式将给出多元正态后验分布（扩展了一个世纪前拉普拉斯的结果），尽管皮尔逊和菲隆警告不要对倾斜的频率分布这样做。现代读者会认出他们得到的分布就是有时用来近似最大似然估计分布的正态分布，但 Pearson 和 Filon 在估计的选择上没有做出这样的限制，而是不顾一切地将其应用于各种估计，特别是矩法估计。

事后看来，结果可能一团糟，甚至不适用于所举的例子，而且这种方法很快就被皮尔逊本人抛弃了。但它为双变量正态相关系数得出了一些正确的结果，而且它大胆而肯定对罗纳德·费舍尔这样的读者具有很大的启发性，我现在要介绍他。我最近发表了一项详细的研究（Stigler，2005），研究了费舍尔是如何在 1922 年撰写他的分水岭著作《理论统计学的数学基础》的，所以我只简要回顾一下那本回忆录的要点。

4. R · A · 费舍尔

在剑桥，费舍尔研究了误差理论，甚至在 1912 年发表了一篇短文，赞扬了高斯估计方法的优点，特别是正态分布样本的标准差。他对由此得出的估计值的不变性非常着迷，例如，频率常数平方的估计值就是常数估计值的平方，因此他称该标准为“绝对的”（费舍尔，1912 年）。但他当时的方法在大多数方面都很肤浅，他默认高斯使用的朴素贝叶斯方法，没有注意到他所考虑的例子中潜在的不一致性，即基于数据分布的标准差平方估计值，即 $\frac{1}{n}\sum(x_i-\bar{x})^2$，与仅将同一原理应用于 $\frac{1}{n}\sum(x_i-\bar{x})^2$ 分布时发现的结果不一致。

四年后，Fisher 向 Pearson 寄送了一篇简短、肤浅的评论，希望能够发表，该评论针对的是 Kirstine Smith 在 Biometrika 上发表的一篇文章，该文章主张使用最小卡方估计方法（Smith，1916 年）。皮尔逊给费舍尔的一封深思熟虑的拒绝信重点指出，选择常数最大化频率函数的方法缺乏清晰和令人信服的理由，皮尔逊甚至表示，他现在认为皮尔逊-菲隆的论文在这方面存在疏忽。他特别提到了史密斯论文中一个敏锐的脚注，该脚注反驳了高斯方法：最大化的概率不是概率，而是概率密度，一个无穷小的概率，如此微不足道的证据有什么力量来捍卫选择？至少最小卡方法是针对实际度量进行优化的。又过了两年，1918 年，费舍尔在估计正态标准差的背景下发现了充分性（Fisher，1920 年，他回忆起皮尔逊提出的提出该方法基本原理的挑战，于是他立即投入了其中，并迅速着手撰写一篇关于统计理论的里程碑式论文，并于 1921 年 11 月向皇家学会宣读了这篇论文，并于 1922 年发表。

5. 费舍尔的第一个证明

根据我的重构，Fisher 在 1922 年那篇伟大的论文中提出了一个简短的论证，随后他发现了充分性；事实上，这是论文中的第一个数学论证。用现代符号表示的论证的本质如下。假设你有两个候选参数 θ 的估计值，分别表示为 S 和 T。假设 T 是 θ 的充分统计量。由于通常情况下，S 和 T 在大样本下都近似为正态分布，让我们（预见到 Wald 将在 1943 年严格提出的一种论证）跟随 Fisher 考虑 S 和 T 实际上具有二元正态分布，$expectation = \theta$，标准差为 σ_S 和 σ_T，相关性为 ρ。然后，二元

正态分布的标准事实告诉我们，$E(S|T=t)=\theta+\rho(\sigma_S/\sigma_T)(t-\theta)$。由于 T 是充分的，因此这不能依赖于 θ，这只有在 $\rho(\sigma_S/\sigma_T)=1$ 或 $\sigma_T=\rho\sigma_S\leq\sigma_S$ 时才有可能。因此，T 的均方误差不可能大于任何其他此类估计值 S，因此根据明确的度量标准（预期平方误差），它必须是最优的！Fisher 一下子就得出了（如果接受用精确正态性代替近似正态性）简单而有力的结果：

> 充分性意味着优化性，至少与一致性和渐近正态性相结合时。

问题是，这个结果有多普遍？费舍尔和后人都没有将一致性和渐近正态性视为主要限制。毕竟，谁会使用不一致的估计，虽然有明显的例外，但渐近正态性不是普遍规则吗？事实上，费舍尔清楚地知道结果比这更强，充分估计以更强的意义捕获了数据中的所有信息；论点只是用一个特定的标准，即最小标准误差来提出主张。但充分性呢？

此时，Fisher 似乎犯了一个有趣且非常有效的错误。他很快研究了许多其他参数示例，并得出结论：最大化似然总是会导致估计值成为充分统计数据的函数！当他于 1921 年 11 月向皇家学会宣读这篇论文时，他的摘要（刊登在《自然》(1921 年 11 月 24 日)）强调道：“通过最大似然法获得的统计数据始终是充分统计数据。”由此可以得出，除了可能需要一致性和渐近正态性这一小点争议外，最大似然估计值始终是最佳的。经过一个半世纪的酝酿，一个真正美丽的理论诞生了。

就在论文准备发表时，最能理解这一理论的人——费舍尔本人——也开始产生怀疑。摘要中大胆的主张并没有出现在发表的版本中，也没有出现对其的否认。他是这样表达的：

> “为了解决估计问题，我们需要一种方法，该方法对于每个特定问题都能自动引导我们得出满足充分性标准的统计数据。我相信，最大似然法提供了这种方法，尽管我对我为此提出的任何证明的数学严谨性都不满意。请阅读下文的读者就最大似然法在任何情况下是否会导致统计数据不足形成自己的观点。就我而言，我很乐意推迟出版，直到能够制定出严格完整的证明；但该方法揭示的新结果的数量和种类迫切需要出版，同时我并非没有意识到纯数学家的合作给应用数学带来的好处，而这种合作往往是由应用数学作者的不完美之处引起的” (Fisher, 1922, page 323)。

除了我上面提到的瓦尔德式论证外，1922 年的论文还提出了几个相关论证。它不那么大胆地陈述了 1921 年摘要中的陈述的反面，即 “似乎任何满足充分性条件的统计数据都必须是通过最优方法 [例如最大似然] 获得的解决方案”（第 331 页）。但费舍尔现在并没有声称充分的统计数据总是存在的。相反，费舍尔给出了一个改进的非贝叶斯版本的皮尔逊-菲隆渐近正态性论证，扩展了关于真实值的似然函数，并指出了该论证如何以及为什么需要最大似然估计（并且它不适用于矩估计），以及如何使用它来评估最大似然估计的准确性（第 328-329 页）。在那里，在一个长脚注中，他指责卡尔皮尔逊没有早点注意到 1898 年论文中的错误。费舍尔指出，皮尔逊在 1903 年发表了矩估计的正确标准误差，尽管他引用了 1898 年的论文，但没有指出 1898 年给出的几个例子的标准误差是错误的。在 1922 年的论文中，费舍尔还尖锐地包含了一个部分，说明了皮尔逊 III 型分布（伽马分布）的最大似然法的用法，将他的结果与皮尔逊和菲隆在 1898 年为同一家族给出的错误结果进行了对比。

6. 三年后

到 1925 年，费舍尔早先的乐观情绪有所消退，他准备了修订版的理论，提交给剑桥哲学学会。在此期间的某个时候，他意识到并非总是存在与参数相同维度的充分统计数据。是什么让他意识到这一点的？费舍尔没有说，尽管他在 1935 年的一次讨论中写道：“我应该提到，如果存在充分统计数据，那么它可以通过最大似然法给出，这一定理已在我 [1922] 的论文中得到证明……。正是这一点让我特别重视这种方法。然而，我当时并没有意识到没有充分统计数据的情况，也没有意识到除了最大值的位置之外，似然函数的其他属性可以弥补不足” (Fisher, 1935, page 82)。我推测他是在考虑一个不存在充分统计数据的问题时学到这一点的，即 1925 年论文中突出出现的问题，即柯西分布的位置参数估计。无论如何，在那篇 1925 年的论文中，费舍尔并没有详细阐述这一不足之处；恰恰相反。论文第 14 页只是随意地提到了不需要存在充分统计数据的可能性，而 1922 年和 1925 年的报纸的读者甚至可能没有注意到发生的细微的重点转变。

1922 年，费舍尔从一致性和充分性入手，1925 年，他从效率入手。在撰写一致性和渐近正态估计时，他指出：“效率标准要求，一个统计量（我们讨论的类别）的方差乘以 n 趋向的固定值应尽可能小。满足此标准的统计量即为有效统计量”（第 703 页）。考虑到这一点，他现在的主要主张是（第 707 页），“我们将看到，最大似然法将始终提供一个统计量，如果该统计量在大样本中呈正态分布，且方差与样本数量成反比下降，则该统计量将是一个有效统计量。”

因此，1925 年，该理论认为，如果存在有效统计数据，则最大似然估计就是有效的。当存在充分且一致的估计时，它也将是最大似然，但这对于效率而言并非必要。他承认可以存在多个有效估计，但他重复了 1924 年已经给出的证明（Fisher，1924a）任何两个有效估计值都具有相关性，随着 n 的增加，相关性接近 1.0。

7. 1925 年“方差分析”证明

费舍尔如何证明这一新的基于效率的公式？他在 1922 年的处理主要依赖于充分性，但这种充分性已不再普遍可用。取而代之的是，他依靠一种新的、有限但数学上相当巧妙的证明，我将其称为“方差分析证明”。该证明显然是基于方差分析平方和分解的概率版本，费舍尔大约在同一时间为农业田间试验单独开发了该版本。费舍尔自己在 1925 年提出的论证相当晦涩难懂，并没有通过考虑逻辑来直接解释；1935 年，他给出了 **1935**, **492**-(44) 的理论细节的改进版本。证明的数学细节显然已被 Hinkey (**1980**) 重新进行了详细阐述。我将满足于仅提供一个强调论证本质的概述，我需要将费舍尔在 1925 年论证中的逻辑发展分为两部分，就像费舍尔在 1935 年版本中所做的那样。

令 $f(x;\theta)$ 为单个观测的密度，令 ϕ 为 n 个独立观测样本的似然函数，即 $\log\phi = \Sigma \log f$。按照 Fisher 的说法，令 $X = \frac{1}{\phi}\frac{\partial\phi}{\partial\theta} = \frac{\partial}{\partial\theta}\log\phi$，我们现在有时称之为得分函数。Fisher 在这里只关心可以通过解方程 $\mathrm{X} = 0$ 求　找到最大似然估计的情况。该论证的第一部分实际上更像是对他在 1922 年所展示内容的重述：通过展开泰勒级数中的得分函数，他得出得分函数近似于最大似然估计的线性函数；正如他所说，$X = -nA(\theta - \hat{\theta})$“如果 $\theta - \hat{\theta}$ 是 $n^{-1/2}$ 阶的小量”，其中他的 $-nA$ 表示我们现在称为样本中的 Fisher 信息，$I(\theta)$。由于在相当普遍的规律性条件下，$E(X) = \int \frac{1}{\phi}\frac{\partial\phi}{\partial\theta}\phi = \int \frac{\partial\phi}{\partial\theta} = \frac{\partial}{\partial\theta}\int\phi = \frac{\partial}{\partial\theta}1 = 0$，我们也有 $\mathrm{Var}(X) = I(\theta)$。正如 Fisher 所指出的，$I(\theta)$ 可以从任何替代表达式中找到

$$I(\theta) = -E\left(\frac{\partial^2 \log \phi}{\partial \theta^2}\right) = E\left(\frac{\partial \log \phi}{\partial \theta}\right)^2$$
$$= -nE\left(\frac{\partial^2 \log f}{\partial \theta^2}\right) = nE\left(\frac{\partial \log f}{\partial \theta}\right)^2$$

费舍尔没有讨论线性近似在何种条件下是充分的；他满足于将其作为最大似然估计的渐近分布的简单途径，即 $N(\theta, 1/I(\theta))$。到目前为止，他还没有超越 1922 年的论证。

1925 年的论证部分是新颖的，即“方差分析证明”，其内容如下：假设 T 为 θ 的任意估计值，假设其一致且渐近正态 $N(\theta, V)$。在证明中，Fisher 将其用作 T 的精确分布，并进一步将 V 视为不依赖于 θ，这大致就是我们现在所说的“常规”参数问题中“合理”估计 T 的情况。Fisher 将得分函数 X 视为样本的函数，并以两种方式查看其在不同样本中的变化。第一种是考虑 X 在所有样本中的总变化，即其方差 $\mathrm{Var}(X) = I(\theta)$。对于第二个，他计算了 $\mathrm{Var}(\hat{X}|T)$，即给定样本的 T 值时 X 的条件变异（即，所有样本中 X 的方差，这些样本的 T 值相同）。由此，他计算出 $E[\mathrm{Var}(X|T)]$，他发现它等于 $Var(X) - 1/V$。由于 $\mathrm{Var}(X) = E[\mathrm{Var}(X|T)] + \mathrm{Var}[E(X|T)]$（这是我提到的类似 ANOVA 的细分），这将得出 $\mathrm{Var}[E(X|T)] = 1/V$。但 $\mathrm{Var}(X|T) \geq 0$ 总是如此，这意味着必然 $E[\mathrm{Var}(X|T)] \geq 0$，因此 $\mathrm{Var}(X) - 1/V \geq 0$。这给出了 $\frac{1}{V} \leq I(\theta)$，或 $V \geq \frac{1}{I(\theta)}$，有效估计相等——我们现在称之为信息不等式。因此，如果最大似然估计确实具有渐近方差 $1/I(\theta)$，则他就建立了效率。

证明的逻辑——以及导致 Fisher 得出该证明的可能路径——似乎很清楚。如果存在充分的统计量 S，那么因式分解定理（Fisher 至少在 1922 年部分地认识到了这一点）将给出 $\phi = C \cdot h(S; \theta)$，其中比例因子 C 可能取决于样本但不取决于 θ。充分性，X 将仅通过 S 依赖于样本，因此对于所有 S 值，$\mathrm{Var}(X\widehat{|}S) = 0$，因此 $E[\mathrm{Var}(X|S)] = 0$ 也成立。此外，如果 S 足够，则最大似然估计（通过求解 $X = 0$ 得到 θ）是 S 的函数。T 无法捕获样本中的所有信息，这反映在给定 T 时 X 值的变化中，即通过 $\mathrm{Var}(X|T)$ 和因此 $E[\mathrm{Var}(X|T)]$。后者起着残差平方和的作用，衡量了 T 相对于 S 的效率损失（或者至少是相对于在有足够统计数据的情况下可以实现的效率损失）。

更重要的是，这种解释为费舍尔提供了一个目标，让他尝试衡量丢失的信息量，甚至确定如何恢复信息，就像在方差分析中，人们可以通过引入导致残差平方和减少的因素来推进分析一样。在 1925 年的论文的其余部分，费舍尔正是在进行这样的研究。他引入了辅助统计量这个术语和概念，实际上是一个协变量，旨在将残差平方和降低到理论上可实现的最小值。他特别关注多项式问题，并专注于研究没有足够估计时的信息损失，以及使用有效但不是最大似然的估计（例如，最小卡方估计）时的信息损失。他发现后者的差异趋向于有限的极限，这是 C. R. Rao（1961 年、1962 年）后来称之为“二阶效率”的度量。

到 1935 年，Fisher 显然已经意识到论证的第一部分（即确定最大似然估计实际上达到了下限 $1/I(\theta)$ 的部分）并不令人满意，因此他提出了一个不同的论证来代替它，以表明已经达到了下限。该论证（Fisher，1935 年，第 45-46 页）源自我将称之为他的第三个证明；我将在后面对此进行评论。

费舍尔 1925 年的作品在概念上非常深刻，并成为许多富有成果的现代讨论的主题，特别是埃夫隆 (1975、1978、1982、1998)、埃夫隆和欣克利 (1978) 以及欣克利 (1980)。

8. 1925 年之后：与 Hotelling 的通信

费舍尔的漂亮理论变得更加复杂，但仍然非常有吸引力。费舍尔在 1925 年提供的证明既不能满足他在 1922 年提到的纯数学家，也无法经受住四分之一世纪后的挑战。这些证明就是他所能提供的全部吗？要回答这个问题，听听费舍尔与一个不怀好意、智力超群的人之间的对话会有所帮助。英国的许多听众对这个问题感兴趣，他们有自己的私人目的，而费舍尔对卡尔·皮尔逊的直言不讳的嘲讽，尽管是以合理地指出皮尔逊先前工作中的重大错误的形式出现的，也只是激怒了他们。但有一位读者在数学方面与费舍尔相近，而且在地理上（他在加利福尼亚）和科学上（他当时正在研究作物估价）都与他相距甚远，以至于他能够参与这样的对话。我指的是哈罗德·霍特林。

1924 年，霍特林以点集拓扑学的论文获得普林斯顿大学博士学位。同年，他加入了斯坦福大学食品研究所，在那里研究农业问题。不久之后，他通过费舍尔 1925 年出版的《研究人员的统计方法》一书发现了费舍尔。霍特林为 JASA 审阅了这本书；事实上，他审阅了前七版，其中前三版是自愿审阅的，不是编辑要求的（Hotelling，1951）。他开始与费舍尔通信，并试图让费舍尔在 1928 年和 1929 年访问斯坦福，但没有成功（Stigler，1999a）。

经过几次友好的信件往来后，1928 年 10 月 15 日，费舍尔（他曾多次收到其他人要求他提供详细数学证明的请求）写信询问霍特林：“现在，我想听听你对收集这些理论碎片的实用性的看法，因为这些碎片是证明我的实际方法所需要的。”霍特林于 12 月 8 日回复，强烈鼓励这项工作，因为它对整个数学都很有价值，并表示“了解相信理论的理由有助于消除甚至围绕合理学说而出现的荒谬观念。”费舍尔在 1928 年圣诞前夜的回复中提议他们合作：

> 28 年 12 月 24 日
>
> 亲爱的 Hotelling 教授
>
> 你的信在圣诞前夕寄到，让我在节日期间有了许多思考。你不会对我的回信抱有太多的期待，因为你看得出我先写信，然后再思考；但我已经明白我有很多事情要感谢你。
>
> 经过几个小时的考虑，我认为正确的做法是向您发送一份内容草稿，您可以随意将其拆分或重写，并告诉您，如果您愿意成为合著者并负责纯数学部分，我将尽力满足您的要求。如果您同意这一点，并像编辑一样，首先做出是否收录或删减的决定，并明确了解我们任何一方一旦认为不值得就可能将其丢弃，我将开始发送内容。它将大部分是新的，因为许多证明可以做得比我以前的出版物好得多。
>
> 您有我所有的旧东西吗？我相信您有，但如果没有，我会尝试找到仍然缺少的东西。
>
> 这似乎是一项艰巨的工作，但如果我不需要过多考虑安排，我就不会抱怨。
>
> 此致
>
> R.A. 费舍尔 [Hotelling 论文集 3]

Fisher 的草稿目录列于下面的附录 1 中。这项工作从未完成。两人之间没有明显的分歧，但随着项目的进行，Fisher 越来越关注遗传学，因为他的 1930 年著作《自然选择的遗传理论》在媒体上发表，而 Hotelling 于 1931 年转入哥伦比亚大学经济学系，这可能是兴趣下降的原因。到 1930 年 2 月，

Fisher 写道：“以教科书的方式完成任何严肃的工作都是一项艰苦的工作；不过，我希望你能坚持你的工作；以及发展纯数学的发展。” 尽管如此，Hotelling 在 1929 年下半年在 Rothamsted 呆了近六个月，在此期间见到了 Fisher 很多次。Hotelling 于 12 月下旬返回美国，及时向美国数学学会 (AMS) 提交了一篇论文，并在 12 月 31 日在得梅因举行的会议上进行了展示。该论文的标题是 “最佳统计数据的一致性和最终分布”；即关于最大似然估计的一致性和渐近正态性。它发表在 1930 年 10 月的《美国数学会志》上。

可以合理地猜测，该论文所采用的方法在一定程度上反映了费舍尔的观点，这是在与费舍尔长期会面后直接提出的，尽管费舍尔显然没有直接参与写作。无论如何，当费舍尔于 1930 年 1 月 7 日写信给霍特林感谢他提供一份副本时，费舍尔唯一的抱怨是霍特林给出的 “一致性” 定义与费舍尔略有不同。费舍尔写道，

> “为了避免将来产生混淆，值得注意的是，你对一致性的定义与我的定义有些不同。在我看来，如果一个统计数据随着样本的无限增加而趋向于错误的极限，那么它就是不一致的。我认为我从未试图将一致性或不一致的区别应用于趋向于无极限的统计数据，而你却称它们都是不一致的。因此，我不应该将样本的平均值称为
>
> $$\frac{1}{\pi}\frac{dx}{1+(x-m)^2}$$
>
> 一个不一致的统计数据，尽管你会这样认为。祝贺你发表了一篇非常优秀的论文。

如今很少有人提及霍特林的论文，这似乎很可惜。霍特林的论文写得非常漂亮，霍特林的大部分作品也是如此，除其他外，他还比费舍尔更清楚地解释了费舍尔自己在这个主题上的工作。他回顾了费舍尔对渐近正态性的证明（基于皮尔逊-菲隆方法的证明），并温和地指出 “不清楚哪些条件，特别是连续性的条件，是使给出的证明有效的必要条件。” 为了弥补这一疏漏，霍特林为一个连续变量的情况提供了两个明确的证明，并过于自信地指出 “扩展到任何数量的变量都是非常明显的；离散变量的相应定理立即成立……” 问题是，正如霍特林的清晰阐述向现代读者表明的那样，证明不起作用。他通过反正切变换将参数空间转换为有限区间（如果需要）来简化问题，并通过将观测变量分组为有限数量的小区间来离散化，但没有意识到两者结合起来并不能确保他实现所需结果所需的一致性，除非离散分布具有有界的参数集。1931 年 12 月 5 日，霍特林发布了 37 个 “统计理论中的突出问题” 清单，显然这一错误引起了霍特林的注意。清单上的第 16 个问题是 “证明（霍特林，1930 年）证明中使用的双重限制过程的有效性，尽可能适用于一般情况。”

9. 令人讨厌的小事实的几何阴影

到目前为止，甚至还没有迹象表明未来会出现任何可能玷污这一美丽理论的肮脏丑陋的小事实。但随后，1930 年 11 月 15 日，霍特林写信给费舍尔，提出了一些尖锐的问题。这封信反映了霍特林似乎在 1926 年在费舍尔的作品中发现的推理问题的几何观点，并在他们在罗瑟姆斯特德的谈话后进一步发展。霍特林在 1930 年的论文中给出了这种观点的一个陈述（该论文一定是在罗瑟姆斯特德起

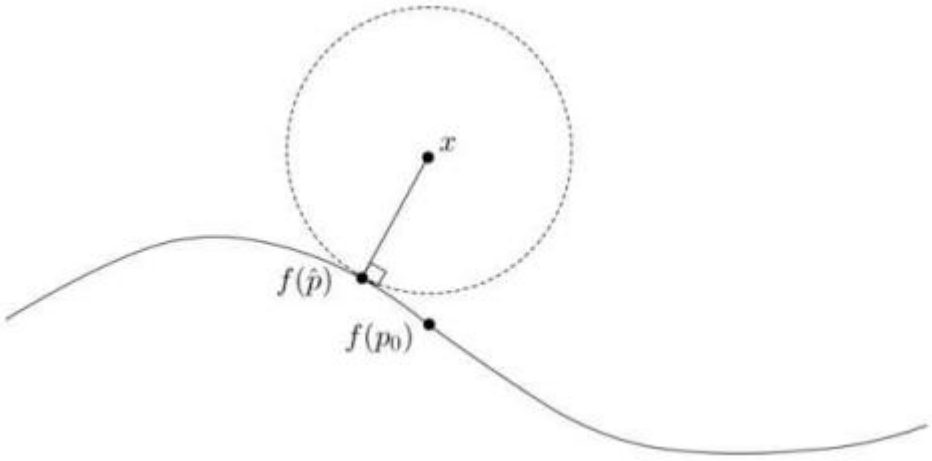

图 2: A reconstruction of Hotelling's geometric view of the multinomial estimation problem, circa Fall 1929. Here x represents a multinomial observed relative frequency vector in the simplex, and the curve f(p) the potential values of the multinomial probability vector; the true value of the parameter p (p0) is shown, as is the MLE and a contour of the likelihood surface.

草的），他在 11 月的信中以不同但等效的符号重述了这一观点。图 2 抓住了其本质，绘制该图是为了展示霍特林用文字和符号传达的内容。

Hotelling 考虑了一个具有 m 个单元格的参数化多项式问题，其中观测值是计数相对频率的向量 $x = (x_1, \dots, x_m)$，取值于 m 维单纯形 $\sum_{t=1}^{m} x_t = 1, x_t \geq 0$ 全部为 t。假设单元格 $f(p) = (f_1(p), \dots, f_m(p))$ 的概率取决于参数 $\hat{p}$；这描述了单纯形中 p 变化时的曲线。假设 $p = p_0$ 表示参数的真实值，假设 p 为 p 的最大似然估计，假设 $f(p_0)$ 和 $f(\hat{p})$ 为曲线上与这两个值相对应的点。在其 1930 年的论文中，Hotelling 进一步指出："似然度 L 在关于 [x] 的近似球面超曲面系统上是常数。点 $[f(\hat{p})]$ 是曲线上与曲线相交的近似球面中最小的一个上的点，因此近似为曲线上距离 [x] 最近的点"(Hotelling, 1930)。

以下是霍特林在通信中提出的问题，其背景一定是他们在罗瑟姆斯特德所采用的共同讨论框架。

> 亲爱的费舍尔博士：
>
> 非常感谢您最近的来信，其中附有图表和数据。
>
> 最近我一直在研究最大似然法中的各种问题；我想知道您是否可以告诉我，在什么条件下，您的证明对通过这种方法获得的统计数据的最小方差有效，或者更确切地说，该定理的确切含义。几个问题之一是，统计数据的方差或其与真实值的均方偏差是否应该用作准确度的度量。
>
> 用 $\hat{p}$ 表示参数 p 的最佳估计，其真实值为 p_0，是否可以说，假设 $\hat{p}$ 服从正态分布，则 $\hat{p}$ 的方差小于相同观测值的任何其他函数的方差？显然，没有进一步的限定，就不能说，因为观测值的函数可以定义为具有任意小的方差。因此，我们必须将比较限制在适合估计 p 的一类特殊函数上，但此类的定义不能涉及 p0。该类应如何定义？作为一致统计类？如

果是这样，就必须面对以下困难。

考虑有限个 m 个类别中的频率分布，涉及参数 p。在 n 个样本中，设 x_t 为属于第 t 个类别的人数 [Hotelling 显然是指相对频率]。设 $f_t(p)$ 为某个人属于此类的概率。如果我们在 m 空间中取 $x_1, \ldots, x_m$ 作为坐标，则方程

$$x_t = f_t(p) \quad (t = 1, \ldots, m)$$

表示以 p 为参数的曲线。与样本相对应的点将形成一个“球状星团”（正如您在 1915 年所说的那样）[2] 关于曲线上 $p = p_0$ 的那个点。对于大样本，最大似然法大致相当于取 $\hat{p}$ 曲线上最接近代表样本的点的参数；即正交投影。现在考虑将样本点投影到曲线上的某种其他方法；例如正交投影，然后沿曲线交替拉伸和收缩。那么，如果 p0 恰好位于稠密区域（即高密度）中的一个点，那么对于足够大的样本，这种估计方法将产生一个方差小于最大似然法的统计量。当然，如果真实值 p0 位于稀疏区域（即低密度），其方差将更大，并且对 p0 的不同可能值进行平均可能会显示比最佳统计量更大的平均方差。但这种平均似乎与“贝叶斯定理”一致，因为它假设先验概率相等。

霍特林接着指出，即使在对称 beta 密度的特殊情况下，最大似然也未达到最优，但他的推导因一个简单的微分错误而受损。在收到 Fisher 1930 年 11 月 28 日的回复之前，霍特林于 12 月 12 日再次写信，纠正了自己关于 beta 估计问题的错误，并扩展了他的其他评论，相当清楚地推测了超高效估计的可能性。

最佳统计数据在何种情况下方差最小这一一般问题极其有趣。从某些具有不连续性的分布的考虑来看，该属性并非完全普遍，这一点似乎很清楚；此外，如果知道真实值，就可以设计出一种估计系统，使其方差任意小；即使真实值未知，也可能采用这种估计系统。

我有两个学生正在研究上述曲线和你处理的 III 型情况的 m 的最佳估计值。由于无法通过纯数学方法对小样本做出任何有意义的结果，他们可能很快就会求助于实验。[3]。

谨致问候，

哈罗德 • 霍特林

霍特林的信以直接但非对抗的方式提出了一个具有挑战性的问题。霍特林说，显然需要对估计类别进行更多限制；他们在洛瑟姆斯特德显然认同的几何观点表明，仅靠一致性是不够的。没有明显的保证表明曲线 $f(p)$ 和似然轮廓使得不可能超过最大似然。需要做些什么来防止这种情况发生，或者至少让像霍特林这样的读者相信这种担心是毫无根据的？霍特林假设的改进当然很模糊。现代读者可

[2] 1Hotelling 显然指的是 Fisher 在 Fisher（1924 年，第 101 页）中使用令人回味的天文学术语“球状星团”来描述点云。Fisher（1924 年）在 Fisher（1915 年）中总结了他在多维空间方法时使用了该术语。Fisher 在 Fisher（1915 年）中没有使用该术语，尽管它在那里也适用。

[3] 从信件中其他地方的评论可以清楚地看出 Hotelling 所说的“实验”是指用骰子或纸牌进行模拟

能会倾向于将它们视为霍奇斯估计的预兆，甚至是通过斯坦因估计的收缩，但即使它们没有达到这一点，它们也向费舍尔提出了明确的挑战。

10. 费舍尔的答复：最大似然法有效性的第三个证明

到 1930 年，费舍尔对持怀疑态度的读者的质疑并不陌生。他对友好人士的质疑的一般反应是清楚地说明他准备说什么，同时避免直接谈及提出的问题。他不会回应批评，更不会承认其有效性，而是直接转向新的、改进的立场，通常不会表明这是最有力的陈述，甚至可能暗示相反，或者至少允许读者猜测。这里的情况就是这样。

费舍尔对霍特林第一封信的回复很简短，但其中包括一个附件 (A)，其中概述了一个阐明费舍尔观点的新证明，以及第二份短信 (B)，纠正了霍特林在区分 $\quad$ 密度时的错误。

> 1930 年 11 月 28 日
>
> 亲爱的 Hotelling，
>
> 我附上了关于您提出的观点的两份说明 A 和 B。第一份说明引入了多项式的一般方差和协方差，并且通过用多重泊松 3 替换多项式，做得更漂亮，但论证可能更清楚。
>
> 这是一封非常简短的信；信的内容在附件中……
>
> 此致，
>
> R. A. Fisher[Hotelling 论文集 45 号]

Fisher 的附录 A 给出了最大似然效率的第三个新证明的草图，该证明采取了不同的攻击点。整个论证（见下文附录 2）优雅、几何，并且我相信在分析所暗示的默认规律性条件下也是正确的，或者至少是可完成的。他采取的几何立场是他和 Hotelling 在 Rothamsted 讨论的立场，仅限于参数多项式分布族的情况。Fisher 确实引入了对估计类 T 的新限制，这显然是回应 Hotelling 的要求，即进一步限制允许的估计类，以排除 Hotelling 所暗示的那种令人讨厌的小事实。正如 Fisher 所说，现在假设所考虑的估计是零次齐次函数 $T = \phi(x_1, \ldots, x_s$，其中 $(x_1, \ldots, x_s)$ 是计数向量。这一点，以及默认的平滑可微性，使他能够获得许多简单的关系，从而得出他总结如下的结论：

> “因此，一致性标准确定了期望线上所有点的 T 值，而与之结合的效率标准则确定了等值统计曲面与该线的交点方向。因此，所有既一致又高效的统计数据都有与该线相切的表面。最大似然法的表面具有这种类型的平面。”

h 次齐次函数 ϕ 是 $\phi(cx, cy, \ldots) = c^h\phi(x, y, \ldots)$，在 Fisher 时代，它们的主要优势在于，如果可微，则它们满足欧拉关系 $x\phi_x + y\phi_y + z\phi_z + \cdots = h\phi(x, y, z, \ldots)$，其中 ϕx 表示 ϕ 关于 x 的偏导数（例如，参见 Courant，1936 年，第 2 卷，第 108-109 页）。在 Fisher 的案例中，零次齐次函数 ϕ 仅是样本相对频率的函数，并且不依赖于样本大小 n。这可能被视为对估计类别的严格限制（Fisher 对此没有评论），但假设可微分性和欧拉关系与 $h = 0$，以及多项式的确切已知协方差，Fisher 可以轻松表达此类中所有估计 T 的渐近方差。他不需要像在之前的证明中那样，诉诸于用正态分布代替近似正

态分布或假设 T 的方差近似为常数所隐含的规律性假设。然后，使用标准拉格朗日方法来最小化该类内一致估计的方差的渐近表达式，并表明得到的方程也是确定最大似然估计的方程，这很容易。

Fisher 后来只是以一种伪装的形式发表了第三个证明，即他假设估计值 T 可以通过一个限制为相对频率的线性函数的估计方程来找到；也就是说，他没有说明从哪里开始，而是直接跳到了欧拉关系。在这种伪装下，没有几何设置和直觉，它出现在他 1935 年的论文中（Fisher，1935 年，第 45-46 页），在那里它提供了一个改进版本的证明，即最大似然估计实现了信息下限。它也出现在 Fisher（1938 年，第 45-46 页）中 1938 年 1 月，应马哈洛诺比斯的邀请，他访问了印度，并撰写了一篇长达 45 页的论文（第 30-32 页）。这篇论文主要从费舍尔的论文中拼凑而成，总结了他当时的观点。在 1956 年出版的书中，他给出了另一个简化版本（显然只是作为例证），仅限于相对频率本身呈线性的估计值 T（Fisher，1956 年，第 145-148 页）。

霍特林在其中扮演着重要的催化剂角色。他帮助费舍尔重新考虑这个问题，并提供了一群非常敏锐的听众，但他本人并没有对最大似然理论做出进一步的贡献。霍特林在 1929 年在罗瑟姆斯特德的近六个月时间里确实写了另一篇相关论文。这是对参数空间微分几何的研究，有时被称为杰弗里斯信息度量，以杰弗里斯 (1946) 命名 (Kass, 1989; Kass and Vos, 1997)。这篇题为 “统计参数空间” 的论文包括 III 型或伽马密度作为一个例子，也一定是霍特林在罗瑟姆斯特德写的。1929 年 12 月 27 日，在霍特林缺席的情况下，奥伊斯坦 • 奥尔在宾夕法尼亚州伯利恒举行的美国数学学会年会上宣读了这篇论文的摘要。只发表了一份摘要，但 Ore 读到的摘要保留了下来（属于 Hotelling 后来的其他笔记的厚文件夹的一部分），并与摘要一起打印在这里（Hotelling，1930a），作为附录 3。

11. 1950 年的情况

总的来说，Fisher 给出了三个最大似然最优性的证明。第一个证明是在 1922 年，基于错误的观点，即最大似然估计总是充分统计量，并且它依赖于将近似正态分布的随机变量视为正态分布。第二个证明是在 1925 年，我称之为方差分析证明。它也需要同样隐含地诉诸规律性，使用正态性代替近似正态性，并假设估计的渐近方差近似为常数，并且似然足够规则以允许评估和处理各种积分。第三个证明是在 1930 年的通信中（后来在 1935 年以估计函数的形式重新表述，失去了几何原点），对分布（假设多项式）和估计（仅相对频率的平滑可微函数，不随样本大小而变化）施加了更严格的限制，但它给出了更令人满意的证明。即使没有填写所有细节，对于所考虑的有限设置来说，这项任务也相当容易。事实上，第三个证明不受 Hotelling 在 1930 年暗示和 Hodges 在 1951 年明确提出的丑陋小事实的影响，但代价是普遍性。尽管如此，多项式分布在离散情况下是人们所希望的普遍性，并且从第三个证明的几何设置中发展出来的直觉至少表面上保证了结果对连续参数族更普遍。

在接下来的几年里，几位数学家意识到，对如此广泛的理论的严格支持程度并不令人满意，并试图填补 Fisher 有意跳过的空白以及他甚至没有意识到的一些空白。早期的主要努力是由 Joseph Doob 完成的（1934，1936）和美国的 Abraham Wald（1943，1949），法国的 Daniel Dugu´e（1937），以及在战争期间在瑞典写作的 Harald Cram´er（1946，1946a）（他读过 Fisher、Doob、Dugu´e 的作品，但显然没有读过 Wald 的作品）。Doob 和 Wald 都与 Hotelling 关系密切；两人均在卡内基奖学金的资助下在哥伦比亚大学与霍特林合作研究该主题，1934-1935 年与杜布合作，1938-1939 年与沃尔德合作。杜布于 1935 年离开哥伦比亚大学前往伊利诺伊大学，但沃尔德留在哥伦比亚大学，在霍

特林休假期间于 1939-1940 年接替霍特林，1946 年霍特林搬到北卡罗来纳州后，沃尔德再次永久担任该职务。

在这些作者中，杜布和杜古遇到了新的困难（正如霍特林在 1930 年的论文中遇到的一样）；杜布被沃尔德温和地纠正，而杜古的失误显然在十年后，即 20 世纪 40 年代中期，伊迪丝 • 穆里尔首次注意到了这一点，她引起了达穆瓦的注意。沃尔德和克莱默的处理是最令人满意的；两者都将严谨程度提升到了新的高度，尽管两者都受到假设条件的复杂性和施加的限制的影响。沃尔德在 1939 年就已发表了关于估计理论的文章，他 1943 年对最大似然估计的渐近充分性的证明可以看作是费舍尔 1922 年证明的一种完成形式。克莱默也坚定地以费舍尔为基础；事实上，他的发展紧跟费舍尔工作的结构，但有严格的论证和明确的条件陈述。克莱默所呈现的大部分内容可以被看作是费舍尔和霍特林所著书籍的实现，尽管没有涉及几何学。

虽然对费舍尔理论的这种反应（即它不是真的，或者至少没有被证实）在某些方面取得了进展，但另一种反应出现了，即声称该理论并不新颖。在这方面，人们的反应就像 17 世纪对威廉 • 哈维 1628 年血液循环演示的反应一样，当时有人否认所声称的现象的真实性，有人主张希波克拉底在 16 世纪的优先权。公元前 400 年（Stigler，1999 年，第 207 页）。卡尔 • 皮尔逊（Karl Pearson）和他阵营中的一些人直到去世都认为费舍尔的最大似然只是高斯方法，只是重新使用而没有明确提及任何贝叶斯基础。这可以归因于对费舍尔所做之事缺乏理解，这种现象甚至困扰着像 G. Udny Yule 这样的一流老统计学家。Yule 1911 年出版的教科书非常出色，尽管经常修订，但从未对费舍尔做过最肤浅的提及（除了他对皮尔逊关于自由度的修正），甚至在 1936 年的第 10 版中也是如此（Yule，1936 年）。

1935 年，另一位更晚近的索赔人的名字被添加到高斯的名字中，当时亚瑟 • 鲍利 (Arthur Bowley) 在为费舍尔 (1935) 表示感谢时，提请人们注意埃奇沃思在 1908-1909 年的工作，该工作与费舍尔的一些工作至少有表面上的相似之处，即三个证明中第二个的信息不等式。

鲍利显然对费舍尔的这项工作只有模糊的理解，他的评论与 15 年后耶日 • 奈曼的评论相比显得温和。奈曼与费舍尔之间的矛盾始于 1934 年，涉及科学和个人问题，后来演变成一场长期的争执。通常，争论处于低水平：费舍尔在最初的分裂之后，除了偶尔的讽刺（通常是隐晦的，不提及奈曼的名字）外，大部分时间都会忽略奈曼，而奈曼通常会淡化费舍尔工作的重要性和原创性，偶尔会发表更详细的抨击文章（Zabell，1992 年；Kruskal，1980 年）。

1937 年，奈曼 (Neyman) 乐于将最大似然的简单思想归功于卡尔 • 皮尔逊 (Karl Pearson)，并引用皮尔逊 (Pearson) 推导的正态相关系数的乘积矩估计作为 “最可能” 值，使用了皮尔逊后来放弃的高斯方法 (Neyman, 1937, 第 345 页)；1938 年，第 132、136 页；Pearson，1896 年，第 262–265 页）。但在 1951 年，奈曼对费舍尔的关注达到了顶峰，他抓住了埃奇沃思的优先权，并将其作为这场争斗中的修辞武器。在对论文集的评论中（Fisher，1950 年），Neyman 重新提起 Bowley 的发现，指责 Fisher 在 “所谓的最大似然估计的有效性” 方面 “不合理地要求优先权”（Neyman，1951 年）。此外，Neyman 写道：“实际上，Edgeworth 和 Fisher 提供的最大似然估计的有效性证明是不准确的，从整体上看，这一断言是错误的。” 这几乎是对虚假事实的错误发现的虚假优先权要求的指控，这将是知识产权纠纷历史上罕见的三重否定。Savage（1976 年）在评论 Fisher 时考虑到了这一评论，“他并不总是以粗鲁的方式成为无可争议的拥护者。” 另一方面，1938 年，Fisher 曾评论过 Neyman 颇具影响力的《数理统计学讲座和会议》（1938 年），这本书中只有少数几处勉强提到 Fisher。费舍尔的评论只有两句话，第一句无伤大雅，第二句则说：“原创材料不足以证明可以出版成书，而琐碎内容又

太多”（费舍尔，1938–1939 年）。

1951 年 6 月，奈曼还写信给《美国统计协会期刊》编辑 W. Allen Wallis，要求该期刊重印 Edgeworth 1908–1909 年的论文，但没有成功（奈曼论文中的信件，班克罗夫特图书馆）。

但基本问题是什么呢？埃奇沃思是否先于费舍尔？如果是，埃奇沃思是否以某种方式影响了费舍尔？我自己的观点与吉米 • 萨维奇（1976 年，第 20 页）的结论基本一致。447–448），尤其是 John Pratt (1976) 从对 Edgeworth 和 Fisher 的详细研究中得出的结论是，虽然 Edgeworth 在这方面的工作确实有价值，但 1951 年对 “不合理的优先权要求” 的指控毫无根据。在一系列强调使用逆概率进行估计的冗长、晦涩和杂乱无章的论文中，Edgeworth 确实包括了他所谓的 “直接方法，摆脱了逆概率的推测性”。回想起来，他发表了这样一段话，最好将其解释为，在非常有限的估计类（基本上是位置参数的 M 估计）中最大化似然值可以得到标准差最小的估计值。他提供的证明（由变分法专家 A. E. H. Love 教授提出）明确基于施瓦茨不等式，与 Fisher 给出的任何证明毫无相似之处。

没有迹象表明埃奇沃思的这部作品曾对费舍尔或任何其他研究该主题的工作者产生过影响。而且散文晦涩难懂——即使以埃奇沃思的标准来看，也异常晦涩难懂——以至于很难相信除了埃奇沃思本人之外，任何当代读者都能认出其结果。即使在后来，要想认出它，读者也需要手头有费舍尔的作品，并且对埃奇沃思有丰富的了解，或者有强烈的历史或个人动机。鲍利在 1928 年准备一份扩展的纪念摘要时，彻底研究了埃奇沃思的作品和表达方式。奈曼既有历史动机，也有个人动机，以及鲍利 1935 年的提示。即使在今天，任何试图了解埃奇沃思从鲍利 1928 年的总结中取得的成就的人（鲍利，1928 年，第 26-28 页）将完全浮现在脑海中，无论文本被困惑多久。这并不是否认，当人们挖掘出 1908-1909 年的原作时，会有一个有限的结果和一丝超越有限结果的理解。埃奇沃思是一位统计科学家，有着异常敏锐和深刻的头脑（Stigler，1986，第 9 章；1999，第 5 章），他在这里的工作进一步证明了这一点。但是，尽管如此，这项工作仍然是一个独立的部分预期——一个暗示，而不是一个实例，即将发生的事情。

埃奇沃思于 1926 年去世，从未对费舍尔发表评论，而费舍尔则一如既往地坚持己见，拒绝在出版物中认真讨论这个问题。他最坦率的私人声明是在两封信中。第一封是 1940 年 2 月 12 日写给莫里斯 • 弗雷切特的信，他在信中将埃奇沃思的陈述描述为与逆概率令人困惑地联系在一起，尽管数学可以与这种方法区分开来。在那封信中，费舍尔用这些话总结了他的观点：“将这种方法与贝叶斯定理联系起来的混淆似乎最初是由高斯造成的，他当然承认它作为一种估计方法的优点，尽管我不知道他是否证明了任何明确的东西。我不知道有任何关于属性、一致性、效率和充分性的明确陈述，这些属性、一致性、效率和充分性可能表征我 1922 年的论文之前估计”（Bennett，1990 年，第 125 页）。第二封信的日期是 1951 年 7 月 2 日，寄给加利福尼亚人霍勒斯 • 格雷 (Horace Gray)，他曾在 1935 年至 1936 年期间与费舍尔在伦敦共事，他写信给费舍尔，提醒他注意奈曼的评论。费舍尔回信说：“从我自己的经验来看，奈曼是一个恶意的捣蛋鬼……我和其他英国统计学家当然早就熟悉了埃奇沃思 1908 年的论文。现在，没有人会不意识到作者非常困惑。就我个人而言，我应该说，他肯定对我后来所展示的东西有所了解。从任何意义上说，他都比我先行，但许多可证实的事实使这一观点变得难以成立”（Bennett，1990 年，第 138-139 页）。

费舍尔列举的事实包括：(i) 埃奇沃思的研究基于逆概率，(ii) 他只关注位置参数，(iii) 他们共同使用的有效估计方差公式来自皮尔逊和菲隆，但未提及该著作中的重大错误。费舍尔指出，由于到 1903 年谢泼德的作品已经表明矩估计的方差与皮尔逊和菲隆给出的方差不同，这给费舍尔提出了疑问：“皮

尔逊和菲隆的方差有任何有效性吗？是否有任何一类估计实际上都具有这些方差？如果是这样，那么一般如何获得这样的估计？但如果埃奇沃思问他们，那他就远远领先于他的时代了。”费舍尔会给予埃奇沃思“一点暗示”，但仅此而已。有些人可能比费舍尔更了解埃奇沃思，但他们是从不同的历史角度来看的。我认为费舍尔在这个问题上并不欠埃奇沃思任何智力上的帮助，这是他自己的损失。如果他花时间和精力去学习埃奇沃思的见解，他可能会走得更远。萨维奇 (1976) 为这种忽视提供了解释，即费舍尔最初认为埃奇沃思的前提很荒谬，后来“因为很难努力追寻不受欢迎的东西。” Neyman 的评论并不是唯一一篇对 Fisher 的工作提出优先问题的评论。在一篇带有倾向性的 1930 年，查尔斯·格罗夫 (Charles Grove) 在评论 Fisher 的《科研人员统计方法》第 3 版时，似乎声称 Fisher 的所有理论都可以在斯堪的纳维亚的蒂勒 (Thiele)、格拉姆 (Gram) 或查理 (Charlier) 的早期著作中找到。格罗夫 (1930) 并没有关注最大似然法，因为他显然认为该法没有依据，而是提出了这样的观点：蒂勒在 1889 年就已在小样本推断方面领先于 Fisher，特别是在使用 k 统计量估计累积量方面，而格拉姆在回归中使用正交多项式方面也做到了这一点。Fisher 在同一出版物中进行了回应，并在给格罗夫的同事阿恩·费舍尔 (Arne Fisher)（丹麦人，似乎是格罗夫评论的发起者）的私人信件中进行了更生动的回应。Fisher 表示，Thiele“对我们现在使用的一些想法的了解并不比 [Karl] Pearson 更深入”（Grove，1930；Fisher，1931；Bennett，1990，第 313 页）。最近，Thiele 的丹麦语译本（Lauritzen，2002）经过了严谨的翻译，并附有评论，可以更好地评估他的杰出工作，然而，他的著作中没有包括对最大似然估计的贡献。

12. 对最大似然法的质疑

在 Hotelling 于 1930 年 11 月 15 日和 12 月 12 日给 Fisher 的探询信之前，似乎没有人提出过最大似然估计可能实际上表现不佳，或者可能通过另一种方法得到显著改善的可能性。Kirstine Smith 和 Karl Pearson 曾在 1916 年质疑过“高斯方法”与最小卡方的相对优点，但两者之间的差异很小；后来两者都被视为渐近有效的估计。在很大程度上，对 Fisher 最大似然的早期保留意见集中在优先级问题（他是否先于他人提出？该方法是否真的有新意？）和实用性问题（与矩法相比，计算是否太难？）。随着 20 世纪 30 年代最大似然被更广泛地采用，对其有效性的证明（能否设计出严格的一般证明？）的关注度不断增加，这不可避免地导致了它何时会失效的问题。最早的明确例子或许是亚伯拉罕·沃尔德 (Abraham Wald) 在 1938 年与耶日·奈曼 (Jerzy Neyman) 的通信中所述。

1938 年春，沃尔德从维也纳移民到美国，当时希特勒吞并奥地利后不久，他接受了加入当时位于科罗拉多斯普林斯的考尔斯经济学研究委员会的邀请。他整个夏天都和考尔斯在一起，然后于 1938 年秋天加入哥伦比亚大学的哈罗德·霍特林。1938 年 9 月 20 日，在他前往哥伦比亚大学前一周，沃尔德写信给奈曼，寄来了一份关于马尔可夫不等式的承诺手稿，但也描述了他遇到的另一个问题。所描述的问题是他在 Wald (1940) 中处理的问题的略微概括，即当观察到直线上的 n 个点但两个坐标都受独立误差的影响时估计直线。然而，沃尔德在写给奈曼的信中包含了他在 1940 年的文章中省略的一句话：“我已经证明了最大似然法会导致参数的错误估计……（即，导致统计数据的随机极限不等于要估计的各个参数的值）。因此，最大似然法不能应用”（奈曼论文，盒子 14，文件夹 28）。沃尔德表示，他已经解决了这个一般估计问题，适用于独立正态分布误差可能方差不等的情况。[4] 9 月 23 日，

[4] 如果将观测点的均值建模为随机样本，则参数的数量不会随着样本大小而增加，并且它们的最大似然估计在温和条件下是一致的；参见

奈曼回复说，他对这个新问题非常感兴趣，“尤其因为它与我自己正在尝试做的事情非常接近”。十年后，奈曼和伊丽莎白 • 斯科特发表了一个简化版的沃尔德例子，并引用了沃尔德 (1940) 的一般性观点，这是几个参数数量不断增加的最大似然估计不一致的例子之一。该版本中直线为 y = x，两个坐标的误差方差相等，被称为奈曼-斯科特例子。它通常表示如下：Xij 是独立的 N(μj, 2)，其中 i = 1,2，j = 1,...,n，在这种情况下， 2 的最大似然估计始终估计正确值的一半（奈曼和斯科特，1948 年）。

1951 年 6 月，就在 Jerzy Neyman 对 Fisher 论文集的评论发表时，伯克利统计实验室在 Neyman 的总体指导下召开了夏季会议。三个研究小组中的一个负责“从超效率和可识别性考虑中产生的一系列问题”。专注于这一主题的小组由 Joseph L. Hodges, Jr.、Lucien Le Cam 和 Agnes Berger 组成。据推测，当时在伯克利担任助理教授的 Hodges 构建了他的示例；无论如何，这项研究很快就取得了足够的进展，Neyman 安排在 1951 年 12 月 29 日星期六在波士顿数理统计研究所会议上举行了一场关于“估计的效率和超效率”主题的会议。该会议上共有四个演讲者，分别是 Jerzy Neyman（“关于估计的渐近效率问题”）、Joe Hodges（“局部超效率”）、Lucien Le Cam（“关于可能实现估计超效率的参数点集”）和 Joseph Berkson（“回归系数的最小二乘和最大似然估计的相对精度”）（生物统计学，1951 年；Littauer 和 Mode，1952 年）。Neyman 和 Hodges 的演讲均未发表；Le Cam 的演讲被发展为他的博士论文并于 1953 年发表。该出版物（Le Cam，1953 年）包括 Hodges 的例子（归功于 Hodges），Le Cam 除其他外还证明了，虽然超效率显然是可能的，但可以实现超效率的参数点集的勒贝格测度为零。

在接下来的十年中，人们发现或设计了许多其他例子。其中，最不费力的是对两个正态分布的五参数混合的估计问题，其中当任一平均参数等于任何观测值时，似然函数会爆炸到无穷大。这个例子和其他几个例子，包括 Bahadur 的一个重要例子，在 Le Cam 中进行了回顾（1990）和 Cox（2006，第 7 章）。Le Cam 推测，正常混合的例子（在 20 世纪 50 年代的民间传说中为人所知，但当时显然没有发表）是 Jack Kiefer 和 Jacob Wolfowitz 的杰作；Cox（2006，第 134-135 页）认为它在某种程度上是病态的。

这些早期的例子引起了一阵兴奋，但今天大部分并不认为它们会削弱该理论。霍奇斯的例子在首次为人所知时产生了重大影响，但自从勒卡姆的论文发表以来，它被视为一项巧妙但微不足道的技术成就。霍奇斯（见上图 1）表明，你可以局部改进最大似然，基本上是通过将估计值缩小到零，因此它也可以被视为查尔斯 • 斯坦 1955 年的收缩估计的早期暗示，即在多参数问题中可以均匀地改进最大似然。但霍奇斯的例子本身对于有限样本来说，对于参数值不接近零的情况，其效果不如最大似然，而且它并没有被视为一个严重的威胁。沃尔德-内曼-斯科特的例子更具有实际意义，并且仍然可以作为现代高度参数化问题中可能发生的情况的警告，其中数据中的信息可能过于分散而无法实现渐近一致性。正常混合示例至少在计算上仍然具有重要意义，因为它展示了在复杂设置中寻求局部最大值或限制参数空间的必要性。Fisher 从未对任何这些示例发表评论。

在此期间，人们继续进行多次尝试，以完善该理论，给出接近必要和充分条件的严格描述，这些条件描述的是最大似然不会产生误导的情况。随着对该主题的研究变得更加精细和正确，该主题的内在困难也变得更加明显。Wald 和 Cram´er 证明最优性所需的条件列表已经很繁琐，解决方案的基本逻辑也逐渐被人遗忘；事实上，一个问题是，实现严格性有时会导致排除基本示例，例如 Wald (1943) 中对正常标准差的估计。

Kiefer 和 Wolfowitz (1956)。

其后果至今仍可见于最优秀的教科书中，例如 Bickel 和 Doksum (2001) 以及 van der Vaart (1998) 所著的教科书，这些教科书的阐述优雅之处在于策略性地限制了覆盖范围。Bahadur (1964) 给出了一个简洁而优雅的定理，该定理建立在 Le Cam 的工作之上，但只处理了一维参数，并且仅限于参数方差连续的渐近正态估计值。

13. 理论上的错误

在这个故事的许多关头，我们遇到了可能被判定为理论错误的相关工人。也许拉格朗日忽略了曲线，直到最后阶段，他的概率都遵循了曲线，因此可能被判定为错误；这肯定会让他的矩估计方法被费舍尔一代认为是极其低效的。也许高斯使用均匀先验，这使得他的解决方案容易受到参数非线性变换的影响，这应该被视为错误。皮尔逊和菲隆在滥用极限的简单方法时犯了错误，导致它给出了错误的答案（Stigler，2007）。当然，费舍尔的 1921 年提出的充分统计数据始终存在的假设是错误的，而 Hotelling 在 1930 年提出的关于最大似然估计的一致性和渐近正态性的证明不能算作具有所声称的普遍性的正确性。

还有其他一些错误我没有讨论过。当 Lambert (1760) 在一次粗略的陈述中只给出了一个例子时，他得到的答案可以说是错误的。Lambert 的唯一具体结果是针对 $n = 2$ 的，声称在这种情况下样本均值总是给出最可能的结果，而这一说法对于柯西密度来说是不成立的。参见 Stigler (1999, 第 16 章)。后来，Doob 和 Dugu´e 试图纠正 Fisher 的一些疏忽，但出现了更小、更微妙的严谨性失误。但我并不意味着这些先驱者有致命弱点。恰恰相反。如果没有拉格朗日的错误，他可能就不会在那么早的时候发现拉普拉斯变换。如果没有皮尔逊和菲隆，费舍尔可能就不会走上他现在的道路。如果没有费舍尔 1921 年的错误结论，他可能就不会急于完成他的理论，尽管这个理论有缺陷和不完整，但它对 20 世纪理论统计学的推出起到了重要作用。在未知领域进行伟大的探索似乎需要极大的勇气，即使是偶然的失误也能带来重大进展。

14. 结论

尽管存在所有这些困难，最大似然法仍然是现代统计学中最常用和最有用的技术之一。面对 20 世纪 50 年代发现的令人讨厌的小事实，这怎么可能呢？首先，在很多问题中都有坚实的数学支持。Fisher 的证明都可以被辩护为正确，至少如果人们接受显然隐含的规律性条件和假设，包括 1922 年对充分估计的限制，以及 1925 年对最大似然估计线性近似的函数的评分。当然，这种辩护有点同义反复：任何陈述都是真的，如果假设了其真实性所需的所有条件；甚至 Pearson-Filon 推导的 “频率常数的可能误差” 也可能被如此辩护。但这两种情况之间存在很大差异。Fisher 的隐含假设在某种程度上相当清楚（例如平滑可微分性、一致估计），并且 Fisher 本人也很清楚；如果他错误地应用了该理论，我不知道。另一方面，皮尔逊-菲隆的情况则有所不同，正如同一篇论文中不恰当的应用所表明的那样。尽管如此，1938 年之后，尤其是 1950 年代的密集数学研究揭示了费舍尔未曾考虑的潜在问题，即参数数量不断增加、似然函数无界以及局部改进超过最大似然的可能性。费舍尔肯定知道其中一些问题，至少在这些问题发表时，如果不是之前的话。对于第一个问题，他本可以反驳，指出在这种情况下，数据中的信息量（测量是他的开创性进展之一）分布在一个维度空间中，维度空间的增加与样本大小成

比例，因此当然可以预料到一致性问题。但他没有。霍特林很早就向他提出了第三种可能性，即局部改进，但在这里，费舍尔也保持沉默，就像他面对其他例子一样。具有讽刺意味的是，费舍尔本人在 1933 年 1 月 14 日写给埃贡 • 皮尔逊的信中可能解释了这种沉默，在信中，他同情皮尔逊与父亲卡尔相处时遇到的困难："许多有才华的男人因此不愿接受，而这个缺点是年龄无法治愈的"（费舍尔论文）。

个性在这一发展中发挥了一定作用。

费舍尔与奈曼的敌意无疑增加了他对公开讨论尚存问题的领域的顽固抵制，也无疑促使奈曼更加热衷于发现并公开讨论这些问题。后者可能被视为争执的好处：20 世纪 30 年代初，和平时期，只有费舍尔、霍特林和那些受霍特林启发研究这个问题的人（杜布、沃尔德）或非战斗人员（杜古），证明中的问题和理论的局限性并未公开。事实上，直到今天，还没有发表过批评文章明确指出霍特林、杜布或杜古证明中的错误来源；要么早期的作品被忽视，要么仅仅被引用，要么以礼貌的暗示提及，例如证明 "不严格"（例如，Doob，1934；Le Cam，1953）。读者不知道理论的真正问题可能在哪里以及如何出现。敌意滋生了不文明的言论；它也导致了原则性的关注。

然而，尽管存在这些问题，最大似然法还是一次又一次地被证明是有用的，即使在没有普遍定理来支持其使用的情况下也是如此。也许正如费舍尔强大的几何直觉所预见的那样，最大似然法的有效应用范围超出了任何合理可实现的证明，尽管这可能以无意中误入不适用领域为代价。我们现在比费舍尔更了解最大似然法的局限性，但还远远不足以保证在最需要它的复杂情况下安全应用。最大似然法仍然是一个真正美丽的理论，即使悲剧可能潜伏在某个角落。

附录 1：Fisher 1928 年 12 月的目录草案

[Hotelling Papers, Box 3]

一、分布种类和变量分布类型（a）不连续、阶跃积分（b）连续可微积分（c）一般类型，积分不可微但频率不限于零测量通过矩指定特征功能， eitxf(x)dx 或者 eitx dF(x) 它的对数，累积性质累积矩函数或半变量 [原文如此] 说明性案例，正态分布的唯一性，多项式和多重泊松 II. 正态分布 2 分布是 Sn 1（x2 p) 当 xp 服从关于零的单位方差分布时变换的 = n p=1 cpqxp， n p=1 c2 pq = 1, n p=1 cpqcpq = 0 2 在频率中的应用 t = n¯x 的分布 2 [原文如此]；应用于回归系数；z = 1 2log n2 2 1 n1x2 2。三、相关系数分布、偏相关、复相关、超空间处理 IV. 半变量的矩估计此类估计的简单和多重分布 {伦敦数学学会的论文。如果一年内不会发表} 组合方法五、估算理论（与已经完成的一样，但更多地涉及充分统计）最大似然法贝叶斯定理。逆概率和似然。通过皮尔逊曲线的矩效率来说明。VI. 实验设计（不像研究人员的统计方法），更多信息量的使用 VII. 统计力学；论证清晰不将 x! 作为连续函数当 x 很小时！尚未完成！类似的生物学问题。但还是值得一试。福勒现在已经做了很多，但最速下降法仍然显得非常间接，而且这显然限制了统计论证。

附录 2：Fisher 的附件 A

Fisher 1930 年 11 月 28 日写给 Hotelling 的信中的附件 A。信件是打字的，但公式是手写的，从底部开始的第四个公式（ Θ X1 代表 Φ X1）中明显的印刷错误与原文中的一样 [Hotelling Papers, Box 45]

期望线 x = f( ) 等统计曲面（或区域）T = (x1,…,xs) x = 0 若 为零次齐次。

为了一致性 = (f1,…,fs) 对于大样本，只要不存在高达 n−1/2 阶的偏差，

$$V(T) = \text{平均值}\left(\sum \frac{\partial\phi}{\partial x}\,\delta x\right)^2$$

$$= \sum f\left(1-\frac{b}{x}\right)\left(\frac{\partial\phi}{\partial x}\right)^2 - \sum\sum \frac{ff'}{n}\frac{\partial\phi}{\partial x}\frac{\partial\phi}{\partial x'}$$

对于多项式，其中微分指的是期望点。

区分一致性条件，$d\theta = (\sum \frac{\partial\phi}{\partial x}\frac{\partial f}{\partial\theta})\,d\theta$，或 $\sum \frac{\partial\phi}{\partial x}\frac{\partial f}{\partial\theta} = 1$

为了一致性，任何值 $\frac{\partial\phi}{\partial x}$ 都是可接受的，但要满足这个条件，ν 是自由不变的 x 是可以接受的，但要满足这个条件才能保持一致，因此，我们可以最小化这个条件下的方差表达式，并得到形式为

$$f_1(\theta)\frac{\partial\theta}{\partial x_1} - \frac{f_1}{n}\sum f\frac{\partial\phi}{\partial x} = \lambda\frac{\partial f_1}{\partial\theta}$$

x = 0，因此，对于所有类别

现在，如果 ϕ 在零度 x 上是齐次的，则 $\sum f\frac{\partial\phi}{\partial x} = 0$，因此，对于所有类

$$\frac{\partial\phi}{\partial x} = \frac{\lambda}{f}\frac{\partial f}{\partial\theta}\text{or}$$

$$\frac{\partial\phi}{\partial x} = \frac{1}{f}\frac{\partial f}{\partial\theta}\Big/\sum\frac{1}{f}\Big(\frac{\partial f}{\partial\theta}\Big)^2$$

因此，一致性标准确定了期望线上所有点的 T 值，而与之相结合的效率标准则确定了等值面与期望线相交的方向。线。因此，所有一致且有效的统计数据都有与该线相切的表面。

最大似然的表面具有这种类型的平面。

附录 3：参数空间上的 Hotelling

1929 年 12 月 26 日至 29 日，霍特林曾短暂出席在宾夕法尼亚州伯利恒举行的美国数学学会年会，但他在原定于 12 月 27 日宣读本文之前就离开了。在他缺席的情况下，耶鲁大学的 Oystein Ore 教授宣读了这篇论文，Ore 随后将手稿退还给了霍特林；只有摘要被发表（Hotelling，1930a）。

与此同时，霍特林于 12 月 30 日至 31 日前往爱荷华州得梅因参加 AMS 例会，并于 12 月 31 日宣读了他的论文“最优统计量的一致性和最终分布”（霍特林，1930 年）。以下摘要是奥尔宣读的整个手稿，摘自哥伦比亚大学的霍特林论文（第 44 盒）。

统计参数空间

作者为斯坦福大学的 Harold Hotelling。

[摘要] 对于表示频率分布的参数 p1,...,pn 的 n 维空间，统计上显著的度量是通过这些参数的有效估计的方差和协方差来定义的。对于普通类型的分布，这样的空间总是弯曲的。对于正态律的两个参数，流形可以部分表示为具有尖锐圆形边缘的负曲率旋转曲面。在这个曲面上，色散的变化用沿生成器移动来表示。对于任何给定形状的 Pearson III 型曲线 [即伽马分布]，都会出现相同的曲面。对于不受限制的 III 型曲线，有三个参数；研究了它们的空间。给出了统计参数空间通常具有的某些度量属性。

“统计参数空间”摘要

“人口”由函数指定

$$f(x, p_1, \dots, p_k)$$

这样 f dx 就是观测值落在 dx 范围内的概率。在统计理论中，我们给出了观测值 x1,...,xN，并希望估计参数 p1,...,pk 的值。进行这些估计的方法有无数种；但其中一种具有某些特别有价值的特性，那就是最大似然法。似然法定义为

$$\prod_{i=1}^{N} f(x_i, p_1, \dots, p_k)$$

用 L 表示其对数。设 ^p1,..., ^pk 为使 L 最大化的值。R. A. Fisher 将它们称为最优统计数据或参数的最优估计。从 N 个样本得出的估计 ^p − p 的误差分布在 N 较大值时趋近于正态形式

$$Ke^{-\frac{1}{2}T} d\hat{p}_1, \dots, d\hat{p}_k,$$

在哪里

$$T = \sum\sum g_{\alpha\beta}(\hat{p}_\alpha - p_\alpha)(\hat{p}_\beta - p_\beta).$$

这里 $g_{\alpha\beta}$ 是数学期望

$$\frac{\partial^2 L}{\partial p_\alpha \, \partial p_\beta},$$

并且是变换 p 下的二阶协变张量 = (p1,...,pk)——当然，二阶导数本身不是张量。

这个张量性质表明

$$g_{\alpha\beta} \, dp^\alpha \, dp^\beta$$

被视为坐标空间 p1,...,pk 中的距离元素。事实上，相当多的微分几何学可以立即得到新的统计结论。应该立即指出，这些空间不是平坦的，而是以取决于初始人口分布的方式弯曲的。

在各种生物问题中，k 空间中短距离跳跃的“随机迁移”问题都会出现，而进化被认为是通过小突变进行的。此类问题也出现在实验工作中，例如卡特在罗瑟姆斯特德开发的计数土壤细菌的稀释法。对于短距离跳跃，这些问题相当于弯曲空间中的热传导和测地线问题。

如果我们考虑任何固定形状的初始分布曲线，则需要估计两个参数，给出曲线的位置和比例，例如正态误差曲线的平均值和标准差。在这种情况下，我们的 k 空间是恒定负曲率的表面。通过伪球面表示正态曲线，标准差的变化由沿生成器的运动表示，平均值的变化由绕轴的旋转表示。方差越大意味着与轴的接近度越高。

由于伪球面上位于同一子午线上两点之间的测地线比子午线更靠近轴，因此我们得出了一个有趣的生物学结论。如果两个相关物种的方差大致相同，但均值不同，则最有可能的共同祖先的方差比现存物种的方差更大。

对于皮尔逊 III 型曲线，位置和尺度的测量值不是沿测地线变化，而是沿斜航线变化。

统计参数空间可用于处理各种问题，这些问题需要检验假设与观测之间的差异，其中涉及两个或更多个观测。因此，如果要检验的假设是某个物种的频率分布在某个维度上具有正常形式，该物种是由另一个物种的一系列小突变产生的，并且如果我们考虑方差的差异以及均值的差异，我们就会将 $n = 2$ 的 X^2 分布应用于 $n = 2$，就像在判断射击技术时，我们可以将射击与目标中心的垂直偏差与水平偏差结合起来一样。但是，均值和方差坐标所在的表面是伪球面而不是平面，这一事实表明，必须对从 X^2 计算出的更大偏差的概率进行校正。事实上，测地圆的面积或周长大于平面上相同半径的面积或周长。面积的过大衡量了必须应用的校正，以获得更大差异的真实概率。

如果我们用代表任何种群的伪球面上的某个点来描述一个测地圆，那么圆周上的点就代表统计数据，例如均值和方差，这些统计数据可能以相同的可能性从该种群的一个样本中获得。反过来，如果对应于给定的样本，我们固定一个点作为测地圆的中心，那么圆周上的点就代表根据该样本的证据，所有种群都具有同等可能性。

致谢

感谢 Henry Bennett 允许我查阅和引用阿德莱德的 Fisher 论文，感谢 Michael Ryan 允许我引用哥伦比亚大学珍本手稿图书馆的 Harold Hotelling 论文，感谢 Susan Snyder 允许我引用加州大学伯克利分校班克罗夫特图书馆的 Jerzy Neyman 论文（索书号 BANC MSS 84/30C，盒子 14，文件夹 28）。感谢 Peter Bickel、Larry Brown、Bernard Bru、David Cox、Persi Diaconis、Anthony Edwards、Brad Efron、Tim Gregoire、Marc Hallin、Lucien Le Cam、Erich Lehmann、Peter Mc-Cullagh、Edith Mourier 和 Ingram Olkin 在本次调查过程中提供的评论。本文基于 2006 年 8 月 2 日在里约热内卢举行的 IMS 年会上 Lucien Le Cam 纪念演讲中提供的材料。

参考

Norden (1972–1973) 调查了截至 1972 年的文献，并列出了大量参考书目。Hald (1998) 以现代符号详细报告了 Fisher 在统计推断方面发表的著作，以及 Pearson 和 Filon 以及 Edgeworth 撰写的与最大似然相关的著作。Aldrich (1997) 和 Edwards(1997a) 讨论了 Fisher 最早关于最大似然的著作；

我还从 Stigler (2005) 的不同角度描述了这项工作，并在 Stigler (1973, 2001, 2007) 中描述了 Fisher 工作的其他方面。Edwards (1974)、Kendall (1961) 和 Hald (1998, 2007) 是描述最大似然法前身的人之一；有关此内容和其他参考资料的更多详细信息，请参阅 Stigler (1999, Chapter 16)。Savage (1976)、Pratt (1976) 和 Hald (1998、2007) 阐述了 Fisher 的工作与 Edgeworth 的工作之间的关系。有关几何在 Fisher 的估算工作中所起的作用的最佳理解，请参阅 Efron 的 Wald 讲座 (1982)，有关 Fisher 几何方法的优雅而富有洞察力的现代发展以及历史参考，请参阅 Kass 和 Vos(1997)。有关 Hotelling 的生活和工作，请参阅 Arrow 和 Lehmann (2005)、Darnell (1988)、Hotelling (1990) 以及 Smith (1978)；关于 Fisher 的生平，请参见 Box (1978)；有关皮尔逊生平的详情，请参阅 Porter (2004)。Hotelling– 费舍尔的信件现收藏于哥伦比亚大学霍特林收藏馆（珍本手稿图书馆）和阿德莱德大学的费舍尔文件馆；每份文件都包含比寄出信件更完整的信件。奈曼的文件现收藏于加州大学伯克利分校的班克罗夫特图书馆。

Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood 1912–1922. Statist. Sci. 12 162–176. MR1617519 Arrow, K. J. and Lehmann, E. L. (2005). Harold Hotelling 1895–1973. Biographical Memoirs of the National Academy of Sciences 87 3–15.

Bahadur, R. R. (1964). On Fisher's bound for asymptotic variances. Ann. Math. Statist. 35 1545–1552. MR0166867 Bahadur, R. R. (1983). Hodges superefficiency. In Encyclopedia of Statistical Sciences (S. Kotz and N. L. Johnson, eds.) 3 645–646.

Bennett, J. H., ed. (1990). Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher. Clarendon Press, Oxford. MR1076366

Bernoulli, D.(1769). Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda. Manuscript; Bernoulli MSS f.299– 305, University of Basel. English translation in Stigler (1997).

Bernoulli, D. (1778). Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda. Acta Academiae Scientiarum Imperialis Petropolitanae for 1777, pars prior 3–23. Reprinted in Bernoulli (1982). English translation in Kendall (1961) 3–13, reprinted 1970 in Pearson, Egon S. and Kendall, M.G. (eds.), Studies in the History of Statistics and Probability, pp. 157–167. Charles Griffin, London.

Bernoulli, D. (1982). Die Werke von Daniel Bernoulli. Band 2. Analysis. Wahrscheinlichkeitsrechnung. Birkh¨auser, Basel. MR0685593

Bickel, P. J. and Doksum, K. (2001). Mathematical Statistics. Basic Ideas and Selected Topics, 2nd ed. 1. Prentice Hall, Upper Saddle River, NJ. MR0443141

Biometrics (1951). News and Notes. Biometrics 7 449–450. Bowley, A. L. (1928). F. Y. Edgeworth's Contributions to Mathematical Statistics. Royal Statistical Society, London.(Reprinted 1972 by Augustus M. Kelley, Clifton, NJ.)

Box, J. F. (1978). R. A. Fisher. The Life of a Scientist. Wiley, New York. MR0500579

Courant, R. (1936). Differential and Integral Calculus. Nordeman, New York.Cox, D. R. (2006). Principles of Statistical Inference. Cambridge Univ. Press. MR2278763

Cram´er, H. (1946). Mathematical Methods of Statistics. Princeton Univ. Press. MR0016588

Cram´er, H. (1946a). A contribution to the theory of statistical estimation. Skand. Aktuarietidskr. 29 85–94. Reprinted in H. Cram´er, Collected Works 2 948–957. Springer, Berlin (1994). MR0017505

Darnell, A. C. (1988). Harold Hotelling 1895–1973. Statist. Sci. 3 57–62. MR0959716

Doob, J. L. (1934). Probability and statistics. Trans. Amer. Math. Soc. 36 759–775. MR1501765

Doob, J. L. (1936). Statistical estimation. Trans. Amer. Math. Soc. 39 410–421. MR1501855

Dugu´e, D. (1937). Application des propri´et´es de la limite ausens du calcul des probabilit´es a l'´etude de diverse questionsd'estimation. J. l'´Ecole Polytechnique 3e s´erie (n. 4) 305–373.

Edwards, A. W. F. (1974). The history of likelihood. Internat. Statist. Rev. 42 9–15. MR0353514

Edwards, A. W. F. (1997). Three early papers on efficient parametric estimation. Statist. Sci. 12 35–47. MR1466429

Edwards, A. W. F. (1997a). What did Fisher mean by "inverse probability" in 1912–1922? Statist. Sci. 12 177–184. MR1617520

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). Ann. Statist. 3 1189–1242. MR0428531

Efron, B. (1978). The geometry of exponential families. Ann. Statist. 6 362–376. MR0471152

Efron, B. (1982). Maximum likelihood and decision theory (The 1981 Wald Memorial Lectures). Ann. Statist. 10 340– 356. MR0653516

Efron, B. (1998). R. A. Fisher in the 21st century (with discussion). Statist. Sci. 13 95–122. MR1647499

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. Biometrika 65 457–482. MR0521817

Fienberg, S. E. and Hinkley, D. V. eds. (1980). R. A. Fisher: An Appreciation. Springer, New York. MR0578886

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. Messenger of Mathematics 41 155–160; reprinted as Paper 1 in Fisher (1974); reprinted in Edwards(1997).

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10 507–521; reprinted as Paper 4 in Fisher (1974).

Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. Mon. Notices Roy. Astron. Soc. 80 758–770; reprinted as Paper 12 in Fisher (1974).

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philos. Trans. Roy. Soc. London Ser. A 222 309–368; reprinted as Paper 18 in Fisher (1974).

Fisher, R. A. (1922a). On the interpretation of 2 from contingency tables, and the calculation of P. J. Roy. Statist. Soc. 85 87–94; reprinted as Paper 19 in Fisher (1974).

Fisher, R. A. (1924). The Influence of Rainfall on the Yield of Wheat at Rothamsted. Philos. Trans. Roy. Soc. London Ser. B 213 89–142; reprinted as Paper 37 in Fisher (1974).

Fisher, R. A. (1924a). Conditions under which 2 measures the discrepancy between observation and hypothesis. J. Roy. Statist. Soc. 87 442–450; reprinted as Paper 34 in Fisher (1974).

Fisher, R. A. (1925). Theory of statistical estimation. Proc. Cambridge Philos. Soc. 22 700–725; reprinted as Paper 42 in Fisher (1974).

Fisher, R. A. (1931). Letter to the Editor. Amer. Math. Monthly 38 335–338. MR1522291

Fisher, R. A. (1935). The logic of inductive inference. J. Roy. Statist. Soc. 98 39–54; reprinted as Paper 124 in Fisher(1974).

Fisher, R. A. (1938). Statistical Theory of Estimation. Univ. Calcutta.

Fisher, R. A. (1938–1939). Review of "Lectures and Conferences on Mathematical Statistics" by J. Neyman. Science Progress 33 577.

Fisher, R. A. (1950). Contributions to Mathematical Statistics. Wiley, New York.

Fisher, R. A. (1956). Statistical Methods and Scientific Inference. Oliver and Boyd, Edinburgh.

Fisher, R. A. (1974). The Collected Papers of R. A. Fisher U. of Adelaide Press. MR0505093

Galton, F. (1908). Memories of my Life. Methuen, London. Gauss, C. F. (1809). Theoria Motus Corporum Coelestium. Perthes et Besser, Hamburg. Translated, 1857, as Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections, trans. C. H. Davis. Little, Brown; Boston. Reprinted, 1963, Dover, New York.

Grove, C. C. (1930). Review of "Statistical Methods for Research Workers." Amer. Math. Monthly 37 547–550. MR1522136

Hald, A. (1998). A History of Mathematical Statistics from 1750 to 1930. Wiley, New York. MR1619032

Hald, A. (2007). A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713 to 1935. Springer, New York. MR2284212

Hinkley, D. V. (1980). Theory of statistical estimation: The 1925 paper. Pp. 85–94 in Fienberg and Hinkley (1980).

Hotelling, H. (1930). The consistency and ultimate distribution of optimum statistics. Trans. Amer. Math. Soc. 32 847–859. MR1501565

Hotelling, H. (1930a). Spaces of statistical parameters (Abstract). Bull. Amer. Math. Soc. 36 191.

Hotelling, H. (1951). The impact of R. A. Fisher on statistics. J. Amer. Statist. Assoc. 46 35–46.

Hotelling, H. (1990). The Collected Economic Articles of Harold Hotelling. Springer, New York. MR1030045

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. London Ser. A 186 453–461. MR0017504

Kass, R. E. (1989). The geometry of asymptotic inference. Statist. Sci. 4 188–219. MR1015274

Kass, R. E. and Vos, P. W. (1997). Geometrical Foundations of Asymptotic Inference. Wiley, New York. MR1461540

Kendall, M. G. (1961). Daniel Bernoulli on maximum likelihood. Biometrika 48 1–18. Reprinted in 1970 in Pearson, Egon S. and Kendall, M. G. (eds.), Studies in the History of Statistics and Probability. Charles Griffin, London, pages 155–172.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many parameters. Ann. Math. Statist. 27 887–906. MR0086464

Kruskal, W. H. (1980). The significance of Fisher: A review of "R. A. Fisher: The Life of a Scientist" by Joan Fisher Box. J. Amer. Statist. Assoc. 75 1019–1030.

Lagrange, J.-L. (1776). M´emoire sur l'utilit´e de la m´ethodede prendre le milieu entre les r´esultats de plusieurs observations; dans lequel on examine les avantages de cette m´ethode par le calcul d es probabilit´es, & ou l'on resoud differens problemes relatifsa cette mati'ere. Miscellanea Taurinensia 5 167–232. Reprinted in Lagrange (1868) 2 173–236.

Lagrange, J.-L. (1868). Oeuvres de Lagrange, 2. Gauthier- Villars, Paris.

Lambert, J. H. (1760). Photometria, sive de Mensura et Gradibus Luminis, Colorum et Umbrae. Detleffsen, Augsburg. (French translation 1997, L'Harmattan, Paris; English translation 2001, by David L. DiLaura, for The Illuminating Engineering Society of North America). Laplace, P. S. (1774). M´emoire sur la probabilit´e des causes par les ´evenemens. M´emoires de math´ematique et de physique, present´esa l'Acad´emie Royale des Sciences, par divers savans, & l^u dans ses assembl´ees 6 621–656. Translated in Stigler (1986a).

Lauritzen, S. L. (2002). Thiele: Pioneer in Statistics. Oxford Univ. Press. MR2055773

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and relates Bayes estimates. University of California Publications in Statistics 1 277–330. MR0054913

Le Cam, L. (1990). Maximum likelihood: An introduction. Internat. Statist. Rev. 58

153–171 [Previously issued in 1979 by the Statistics Branch of the Department of Mathematics, University of Maryland, as Lecture Notes No. 18].

Littauer, S. B. and Mode, E. B. (1952). Report of the Boston Meeting of the Institute. Ann. Math. Statist. 23 155–159.

Neyman, J. (1937). Outline of a theory of statistical estimation based upon the classical theory of probability. Phil. Trans. Royal Soc. London Ser. A 236 333–380.

Neyman, J. (1938). Lectures and Conferences on Mathematical Statistics (edited by W. Edwards Deming). The Graduate School of the USDA, Washington DC.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. Econometrica 16 1–32. MR0025113

Neyman, J. (1951). Review of R. A. Fisher "Contributions to Mathematical Statistics." The Scientific Monthly 72 406– 408.

Norden, R. H. (1972–1973). A survey of maximum likelihood estimation. Internat. Statist. Rev. 40 329–354, 41 39–58. Pearson, K. (1896). Mathematical contributions to the theory of evolution, III: regression, heredity and panmixia.

Philos. Trans. Roy. Soc. London Ser. A 187 253–318. Reprinted in Karl Pearson's Early Statistical Papers, Cambridge: Cambridge University Press, 1956, pp. 113–178. Pearson, K. and Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. Philos. Trans. Roy. Soc. London Ser. A 191 229–311. Reprinted in Karl

Pearson's Early Statistical Papers, Cambridge: Cambridge University Press, 1956, pp. 179–261.

Porter, T. M. (2004). Karl Pearson: The Scientific Life in a Statistical Age. Princeton Univ. Press. MR2054951 Pratt, J. W. (1976). F. Y. Edgeworth and R. A. Fisher on the efficiency of maximum likelihood estimation. Ann. Statist. 4 501–514. MR0415867

Rao, C. R. (1961). Asymptotic efficiency and limiting information. Proc. Fourth Berkeley Symp. Math. Statist. Probab. 1 531–546. Univ. California Press, Berkeley. MR0133192

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples, with discussion. J. Roy.

Statist. Soc. Ser. B 24 46–72. MR0293766

Savage, L. J. (1976). On rereading R. A. Fisher. Ann. Statist.4 441–500. MR0403889

Sheynin, O. B. (1971). J. H. Lambert's work on probability. Archive for History of Exact Sciences 7 244–256. MR1554145

Smith, K. (1916). On the 'best' values of the constants in frequency distributions. Biometrika 11 262–276.

Smith, W. L. (1978). Harold Hotelling 1985–1973. Ann. Statist. 6 1173–1183. MR0523758

Stigler, S. M. (1973). Laplace, Fisher, and the discovery of the concept of sufficiency. Biometrika 60 439–445. Reprinted in 1977 in Kendall, Maurice G. and Robin L. Plackett, eds., Studies in the History of Statistics and Probability, Vol. 2. Griffin, London, pp. 271–277. MR0326872

Stigler, S. M. (1986). The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard Univ. Press, Cambridge, MA. MR0852410

Stigler, S. M. (1986a). Laplace's 1774 memoir on inverse probability. Statist. Sci. 1 359–378. MR0858515

Stigler, S. M. (1997). Daniel Bernoulli, Leonhard Euler, and Maximum Likelihood. In Festschrift for Lucien LeCam (D. Pollard, E. Torgersen and G. Yang, eds.) 345–367.

Springer, New York. Extensively revised and reprinted as Chapter 16 of Stigler (1999). MR1462957

Stigler, S. M. (1999). Statistics on the Table. Harvard Univ. Press, Cambridge, MA. MR1712969

Stigler, S. M. (1999a). The Foundations of Statistics at Stanford. Amer. Statist. 53 263–266. MR1711551

Stigler, S. M. (2001). Ancillary history. In State of the Art in Probability and Statistics (C. M. de Gunst, C. A. J. Klaassen and A. W. van der Vaart, eds.). IMS Lecture Notes Monogr. Ser. 36 555–567. IMS, Beachwood, OH. MR1836581

Stigler, S. M. (2005). Fisher in 1921. Statist. Sci. 20 32–49. MR2182986

Stigler, S. M. (2007). Karl Pearson's theoretical errors and the advances they inspired. To appear.

van der Vaart, A. W. (1997). Superefficiency. In Festschrift for Lucien Le Cam (D. Pollard, E. Torgersen and G. L. Yang, eds.) 397–410. Springer, New York. MR1462961

van der Vaart, A. W. (1998). Asymptotic Statistics. Cambridge Univ. Press. MR1652247

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. Ann. Math. Statist. 11 284–300. [A summary of the main results of this article, as presented in a talk July 6, 1939, was published pp. 25– 28 in Report of the Fifth Annual Research Conference on Economics and Statistics Held at Colorado Springs July 3 to 28, 1939, Cowles Commission, University of Chicago, 1939.] MR0002739

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Amer. Math. Soc. 54 426–482. MR0012401

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. Ann. Math. Statist. 20 595–601. MR0032169

Yule, G. U. (1936). An Introduction to the Theory of Statistics, 10th ed. Charles Griffin, London. [This was the last edition revised by Yule himself; subsequent revisions from 1937 by

M. G. Kendall were not greatly changed in emphasis.]

Zabell, S. L. (1992). R. A. Fisher and the fiducial argument. Statist. Sci. 7 369–387. Reprinted in 2005 in S. L. Zabell, Symmetry and its Discontents: Essays on the History of Inductive Philosophy. Cambridge Univ. Press. MR1181418

Affiliation:

Stephen M. Stigler[5]
E-mail: stigler@galton.uchicago.edu

翻译：林绪虹[6]
E-mail: linxuhong@yahoo.com

Silkman Statistical Journal http://cookwhy.com/
published by the Funny Project of Silkman Press http://cookwhy.com
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd

[5]Stephen M. Stigler is the Ernest DeWitt Burton Distinguished Service Professor, Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA e-mail: stigler@galton.uchicago.edu.

[6]软件工程师，数学史爱好者，本文基于原作者 2003 年发表于 NETWORK: COMPUTATION IN NEURAL SYSTEMS 的原文翻译，翻译时间 2024 年 4 月）