21 世纪的 R. A. Fisher

1996 年 R. A. Fisher 讲座特邀论文

Bradley Efron*　　翻译：林绪虹†

1998

摘要

费舍尔是 20 世纪统计学中最重要的人物。本演讲探讨了他对现代统计学思想的影响，试图预测 21 世纪的费舍尔主义倾向。费舍尔的哲学特点是贝叶斯和频率论观点之间的一系列精明折衷，并辅以一些在应用问题中特别有用的独特特征。本演讲探讨了几个当前的研究课题，着眼于费舍尔主义的影响或缺乏影响，以及这对未来统计学发展意味着什么。本文以 1996 年的费舍尔演讲为基础，与演讲内容紧密相关。

1. 简介

即使是科学家也需要他们的英雄，而 R. A. Fisher 无疑是 20 世纪统计学的英雄。他的思想主导并改变了我们的领域，其程度甚至连凯撒或亚历山大都羡慕不已。大部分事情发生在 20 世纪的第二个 25 年，但到我接受教育的时候，Fisher 在美国统计学界已沦为一个次要人物，而 Neyman 和 Wald 的影响力则达到了顶峰。

20 世纪末，人们对费舍尔统计学（Fisherian statistics）的兴趣再度高涨，在英国，他的影响力从未减弱过，在美国和统计界的其他地方也是如此。这种复兴在很大程度上被忽视了，因为它隐藏在现代计算方法耀眼光芒的背后。我在这里的主要目标之一是阐明费舍尔对现代统计学的影响。我将主要通过例子来描述费舍尔思想的优势和局限性，最后引出对费舍尔在 21 世纪统计学领域的角色的一些猜测。

接下来的内容基本上是费舍尔在 1966 年 8 月芝加哥联合统计会议上发表的演讲文本。与标准期刊文章相比，演讲形式具有某些优势。首先，演讲旨在让听众在一小时内快速吸收，从而迫使演讲集中在要点而不是技术细节上。口语往往比期刊论文的灰色散文更生动。演讲鼓励更大胆的区分和个人观点，这在书面文章中很容易受到攻击，但我认为对于未来的推测是合适的。换句话说，这将是一幅粗线条的画作，色彩丰富但细节不足。

细心的读者可能会不太看好这些优点。Fisher

*Bradley Efron is Max H. Stein Professor of Humanities and Sciences and Professor of Statistics and Biostatistics, Department of Statistics, Stanford University, Stanford, California 94305-4065 (email: brad@stat.stanford.edu).

†软件工程师，数学历史爱好者，本文于 1998 年 Bradley Efron 出版于 Statistical Science，2024 年翻译）

的数学论证在力量和经济性方面非常出色，而这些优点大部分都缺失了。粗略的笔触有时会掩盖重要的争议领域。大部分论证都是通过例子而不是理论进行的，我自己的作品中的例子发挥了夸张的作用。参考文献很少，而且没有以通常的作者-年份格式标出，而是以注释的形式收集在文本末尾。最严重的是，一小时的限制要求我有点随意地选择主题，为此我专注于费舍尔著作中对我来说最重要的部分，省略了费舍尔影响的整个领域，例如随机化和实验设计。结果更像是一篇个人论文，而不是系统调查。

这是一场关于费舍尔影响的演讲（我现在将提到它），主要不是关于费舍尔本人或他的思想史。对这部作品本身的更深入研究出现在 L. J. Savage 著名的演讲和论文“重读 R. A. 费舍尔”（On rereading R. A. Fisher）中，这是 1971 年的费舍尔讲座（注：原文为 Fisher lecture，待论证具体指待内容），精彩地描述了费舍尔的统计思想，一位领先的贝叶斯主义者进行了更为深入的研究（Savage，1976 年）。由于 John Pratt 的编辑努力，Savage 的演讲在 1976 年的《统计年鉴》中发表，当时他已经过世了。在文章的讨论中，Oscar Kempthorne 称这是他听过的最好的统计学演讲，丘吉尔 • 艾森哈特也这么说。另一个很好的参考资料是 Yates 和 Mather 为 1971 年出版的五卷费舍尔文集所写的介绍。Joan Fisher Box 1978 年的传记《科学家的生活》（The Life of a Scientist，暂无中文译本）中对费舍尔的权威引用。

永远不要与你的英雄见面，这是一条很好的规则。1961 年，当费舍尔在斯坦福医学院演讲时，我无意中遵守了这条规则，当时没有通知统计系。费舍尔强大的个性在这次演讲中消失了，但我希望他的思想的力量没有消失。英雄主义这个词很贴切地描述了费舍尔试图改变统计思维的努力，这些努力对本世纪统计学发展成为科学领域的一支主要力量产生了深远的影响。“下个世纪会怎样？”是标题中隐含的问题，但这个问题我稍后再谈。

2. 统计世纪

尽管标题如此，但演讲的大部分内容都涉及过去和现在。我将首先回顾 20 世纪的统计学，这是我们这个行业取得巨大进步的时代。在 20 世纪，统计思维和方法已成为数十个领域的科学框架，包括教育、农业、经济、生物和医学，并且最近对天文学、地质学和物理学等自然科学的影响越来越大。

换句话说，我们已经从一个鲜为人知的小领域发展成为一个鲜为人知的大领域。大多数人，甚至大多数科学家，仍然对统计学知之甚少，只知道“.05”这个数字有好处，而钟形曲线（bell curve）也许有坏处。但我相信，这种情况在 21 世纪会有所改变，统计方法将被广泛认可为科学思维的核心要素。

1900 年春，卡尔 • 皮尔逊 (Karl Pearson) 发表了著名的 χ^2 论文，为 20 世纪的统计学带来了良好的开端。统计学的发展由二战前的一批知识巨人奠定了基础：奈曼 (Neyman)、皮尔逊夫妇 (Pearsons)、斯图登特 (Student)、柯尔莫哥洛夫 (Kolmogorov)、霍特林 (Hotelling) 和瓦尔德 (Wald)，其中奈曼的工作影响尤为深远。但从我们在本世纪末的观点来看，或者至少从我的角度来看，占主导地位的人物是 R. A. 费舍

尔 (R. A. Fisher)。费舍尔的影响在统计应用方面尤其普遍，但它也贯穿在我们的理论期刊中。鉴于本世纪即将结束，这似乎是评估费舍尔遗产的活力及其未来发展潜力的好时机。

这次演讲的一个更准确但不那么挑衅的标题应该是“费希尔对现代统计学的影响”。我主要会研究一些当前感兴趣的话题，并评估费希尔的思想对它们的影响有多大。演讲的核心部分涉及当前感兴趣的六个研究领域，我认为这些领域在未来几十年将很重要。这也让我有机会谈谈我们可能很快要处理的应用问题类型，以及费希尔统计学是否会对它们有很大帮助。

不过，首先我想简单回顾一下费希尔的思想以及他所回应的思想。评估费雪统计学的重要性的一个困难是很难说清楚它到底是什么。费雪有许多重要思想，其中一些思想，如随机化推理和条件性，是相互矛盾的。这有点像经济学中的马克思、亚当 • 斯密和凯恩斯原来是同一个人。所以我只打算概述费希尔的一些主要主题，并不试图做到完整或哲学上的调和。这次演讲和其余的演讲将非常简短地介绍参考资料和细节，尤其是技术细节，我将尽量完全避免这些细节。

1910 年，也就是 20 岁的 Fisher 发表第一篇论文的两年前，统计学界的伟大思想清单应该包括以下令人印象深刻的清单：贝叶斯定理、最小二乘法、正态分布和中心极限定理、计数数据的二项式和泊松方法、高尔顿相关和回归、多元分布、皮尔逊的 χ^2 和斯图登特 (Student) 的 t。缺少的是这些思想的核心。这份清单是作为临时设备的巧妙集合而存在的。统计学的情况与计算机科学现在面临的情况类似。

用琼 • 费舍尔 • 博克斯（Joan Fisher Box）的话来说，“整个领域就像一个未经开发的考古遗址，它的结构在瓦砾堆上几乎难以察觉，它的宝藏散落在文献中。”

有两个明显的候选者可以提供统计核心：拉普拉斯传统的“客观”贝叶斯统计，即对未知参数使用统一先验，以及以皮尔逊的 χ^2 检验为例的粗略频率论。事实上，皮尔逊正在通过他的皮尔逊分布系统和矩量法开发自己的核心程序。

到 1925 年，费舍尔已经为统计学提供了一个核心，它与拉普拉斯或皮尔逊方案截然不同，也更引人注目。1925 年的这篇伟大论文已经包含了费舍尔估计理论的大部分主要元素：一致性、充分性、可能性、费舍尔信息、效率以及最大似然估计量的渐近最优性。部分缺失的是辅助性，它被提及但直到 1934 年的论文才得到充分发展。

这篇 1925 年的论文甚至包含一个引人入胜且仍存在争议的部分，即 Rao 所称的最大似然估计 (MLE) 的二阶效率。Fisher 从未真正满足于渐近结果，他说，在小样本中，MLE 丢失的信息比竞争的渐近有效估计量要少，并暗示这有助于解决小样本推断的问题（此时 Savage 想知道为什么人们应该关心点估计量中的信息量）。

费舍尔的伟大成就是为统计估计提供了一个最优标准——衡量任何给定估计问题中可能达到的最佳结果的标准。此外，他还提供了一种实用方法，即最大似然法，即使在小样本中，该方法也能相当可靠地产生接近理想最优值的估计量。

最优性结果是科学成熟的标志。我认为 1925 年是统计理论成熟的一年，在这一年，统计学从临时收集的巧妙技术发展成为一门连贯的学科。统计学在 20 世纪初很幸运地获得了费舍尔。我

们迫切需要另一位费舍尔来开启 21 世纪，正如演讲结束时将讨论的那样。

3. 统计推断的逻辑

费舍尔认为，一定存在一种归纳推理逻辑，可以对任何统计问题得出正确的答案，就像普通逻辑解决演绎问题一样。通过使用这种归纳逻辑，统计学家将摆脱贝叶斯学派的先验假设。

费舍尔的主要策略是将给定的推理问题（有时是非常复杂的问题）逻辑地简化为简单的形式，让每个人都同意答案是显而易见的。他最喜欢的“显而易见”的目标是我们观察到一个正态分布的量 x，其期望值 θ 未知，

$$x \sim N\left(\theta, \sigma^2\right) \tag{1}$$

方差 σ^2 已知。费舍尔说，每个人都同意，在这种情况下，最佳估计是 $\hat{\theta} = x$，而 0 的正确 90% 置信区间（使用费舍尔讨厌的术语）是

$$\hat{\theta} \pm 1.645\sigma \tag{2}$$

费舍尔的归纳逻辑可以称为类型理论，其中问题被简化为一小类显而易见的情况。这在统计学中曾经尝试过，皮尔逊系统就是一个很好的例子，但从来没有如此有力或成功。费舍尔在将问题简化为简单形式（如 (1)）方面非常有才华。他为此发明的一些方法包括充分性、辅助性和条件性、变换、关键方法、几何论证、随机化推理和渐近最大似然理论。自费舍尔时代以来，只有一个主要的简化原则被添加到此列表中，即不变性，而且这个原则现在并不普遍受欢迎。

费舍尔总是偏向于精确的小样本结果，但 MLE 的渐近最优性是他迄今为止最具影响力的，或者至少是最受欢迎的简化原理。1925 年的论文表明，在大样本中，未知参数 θ 的 MLE $\hat{\theta}$ 接近理想形式 (1)，

$$\hat{\theta} \to N\left(\theta, \sigma^2\right)$$

方差 σ^2 由 Fisher 信息和样本大小决定。此外，没有其他“合理”的 θ 估计量具有更小的渐近方差。换句话说，最大似然法会自动生成一个可以合理地称为“最优”的估计量，而无需调用贝叶斯定理。

费舍尔的伟大成就引发了人们对最优性结果的极大兴趣。这一兴趣最引人注目的成果是用于最优假设检验的奈曼-皮尔逊引理，随后奈曼的置信区间理论也随之而来。奈曼-皮尔逊引理对假设检验的作用，就如同费舍尔的 MLE 理论对估计的作用一样，为实现最优性指明了方向。

从哲学角度来看，奈曼-皮尔逊引理与费舍尔方案非常契合：利用数学逻辑，它无需调用贝叶斯先验，就能将复杂的问题简化为显而易见的解决方案。此外，它在应用中非常有用，因此奈曼关于假设检验和置信区间的思想现在在日常应用统计学中发挥着重要作用。

然而，奈曼-皮尔逊引理的成功引发了新的发展，导致了一种更为极端的统计最优性，而费舍尔对此深感怀疑。尽管费舍尔的个人动机在这里值得怀疑，但他的哲学疑虑并非毫无根据。奈曼

的思想后来被沃尔德发展为决策理论，为统计学带来了一种截然不同的精神。

费舍尔的最大似然理论是为了回应上个世纪相当肤浅的拉普拉斯贝叶斯主义而提出的。费舍尔的工作展示了一种更严格的统计推断方法。奈曼-瓦尔德决策理论学派将这种严谨精神发扬光大。对手头问题的严格数学表述（通常措辞非常狭隘）随后给出最佳解决方案成为理想情况。实际结果是一种更复杂的频率论推理形式，具有巨大的数学吸引力。

费舍尔被来自右侧的侧翼攻击打了个措手不及，他抱怨说，奈曼-瓦尔德决策理论家可能准确但不正确。他最喜欢的一个例子是关于中心未知的柯西分布

$$f_\theta(x) = \frac{1}{\pi\left[1 + (x-\theta)^2\right]}. \tag{3}$$

给定 (3) 中的随机样本 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$，决策理论家可能会尝试提供形式为 $\hat{\theta} \pm c$ 的最短区间，该区间以 0.90 的概率覆盖真实 θ。Fisher 的反对意见（在 1934 年关于辅助性的论文中阐明）是，对于不同的样本 $\mathbf{x}$，c 应该有所不同，具体取决于 $\mathbf{x}$ 中的正确信息量。

决策理论运动最终催生了其自身的反改革。由萨维奇和德菲内蒂领导的新贝叶斯学派提出了一种更合乎逻辑、更有说服力的贝叶斯主义，强调主观概率和个人决策。萨维奇-德菲内蒂理论最极端的形式直接否定了费舍尔关于统计推断非个人逻辑的说法。战后，人们对客观贝叶斯理论的兴趣也重新燃起，这种理论本质上是拉普拉斯的，但基于杰弗里斯更复杂的选择客观先验的方法，我将在后面详细介绍。

简而言之，这就是我们在 20 世纪末期面临的三种相互竞争的统计推断哲学：贝叶斯派、奈曼-瓦尔德频率论和费舍尔派。在许多方面，贝叶斯派和频率论哲学彼此对立，而费舍尔的思想则在某种程度上是一种妥协。接下来我想谈谈这种妥协，因为它与费舍尔方法的流行有很大关系。

4. 三种互相竞争的哲学

图 1 中的图表显示了贝叶斯派和频率派之间的四个主要分歧领域。这些分歧不仅仅是哲学上的分歧。我之所以选择这四个类别，是因为它们在数据分析层面上会导致不同的行为。对于每个类别，我都粗略地表明了 Fisher 的偏好立场。

4.1 个人决策与科学推理

贝叶斯理论，特别是 Savage-de Finetti 贝叶斯主义（我在这里重点介绍的那种，虽然稍后我也会谈论 Jeffreys 品牌的客观贝叶斯主义），频率论者（一种贝叶斯主义）强调个体决策者，它在商业等个体决策至关重要的领域最为成功。频率论者的目标是让其推论得到普遍接受。费舍尔认为，统计学的正确领域是科学推论，需要说服整个或至少大多数科学界，你得出了正确的结论。在这里，费舍尔远远偏向了频率论者（从哲学上讲，频率论者的观点准确，但时代错误，因为费舍尔的立场早于萨维奇-德菲内蒂学派和奈曼-瓦尔德学派）。

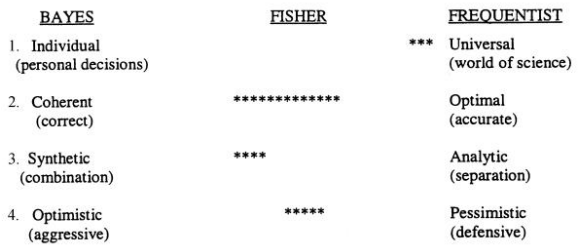

图 1: 贝叶斯方法和频率学派方法之间有四大分歧。我在每一点上都插入了一排星号，粗略地表示了费舍尔推理的首选位置。

4.2 一致性与最优性

贝叶斯理论强调其判断的连贯性，不仅在技术上如此，而且在更广泛的意义上也是如此，即在决策情况的不同方面之间加强一致性关系。频率论意义上的最优性往往是不连贯的。例如，$\exp\{\theta\}$ 的均匀最小方差无偏 (UMVU) 估计不必等于 $\exp\{\theta\}$ 的 UMVU，更严重的是，没有简单的微积分将这两个不同的估计联系起来。费舍尔希望事情既连贯又最优，事实上最大似然估计确实满足

$$\exp\{\hat{\theta}\} = \widehat{\exp\{\theta\}}$$

连贯性和最优性之间的紧张关系就像柯西例子 (3) 中的正确性和准确度之争，其中费舍尔强烈主张正确性。对正确性的强调和对统计推断逻辑存在的信念，使费舍尔哲学向图 1 中的贝叶斯一侧靠拢。费舍尔的实践就不那么清晰了。费舍尔纲领的不同部分彼此并不连贯，在实践中，费舍尔似乎很愿意牺牲逻辑一致性来换取某个问题的巧妙解决方案，例如，在费舍尔信息的频率主义和非频率主义论证之间来回切换。这种针对具体案例的权宜之计是现代数据分析的共同属性，具有频率主义的味道。我将这一类别的费舍尔之星定位得更靠近图 1 的贝叶斯一侧，但分布范围很广。

4.3 综合与分析

贝叶斯决策强调从所有可能的来源收集信息，并将这些信息综合成最终的推论。频率论者倾向于将问题分解成可以单独（并且以最佳方式）分析的独立小块。费舍尔强调使用所有可用信息作为正确推论的标志，在这方面他更赞同贝叶斯的立场。

在这种情况下，费舍尔在理论和方法论上都倾向于贝叶斯立场：基于费舍尔信息的最大似然估计及其附带的近似置信区间理论非常适合来自不同来源的信息的组合。（另一方面，我们有耶茨和马瑟的这句话：“在自己的工作中，费舍尔在面对小型独立数据集时表现最佳……他对收集和分析来自不同来源的与特定问题有关的大量数据从不感兴趣。”他们将此归咎于他在吸烟致癌争议上的固执。在这里和其他地方，我们不得不将费舍尔视为一个失败的费舍尔主义者。）

4.4 乐观与悲观

最后一个类别更多的是心理学而非哲学，但它是扎根于两种相互竞争的哲学的基本性质的心理学。贝叶斯学派在数据分析中往往更具侵略性和冒险精神。选择频率学派哲学的一个极端例子，没有比极小极大理论更悲观和防御性的理论了。它说，如果任何事情都可能出错，它就会出错。当然，极小极大理论的人可能会将贝叶

斯学派的立场描述为“如果任何事情都可能顺利，它就会顺利。”

费舍尔在这里采取了中间立场。他蔑视决策理论家对数学的精细关注（“不仅需要用大炮才能打到麻雀，而且大炮打不到麻雀！”），但他害怕用贝叶斯方法对自然状态进行平均。费舍尔作品真正吸引人的特点之一是其合理妥协的精神，谨慎但不过分关注病理情况。这一直给我留下了对大多数现实问题的正确态度，这无疑是费舍尔在统计应用领域占据主导地位的很大一部分原因。

看看图 1 ，我认为过于努力地从费舍尔的理论中得出连贯的哲学是一个错误。从我们现在的观点来看，它们更容易理解为贝叶斯和频率论思想之间极其精明的妥协的集合。费舍尔通常写作时好像他手头有完整的统计推断逻辑，但这并没有阻止他在想到另一个里程碑式的想法时改变他的系统。

西法雷利（Cifarelli）和雷加齐尼（Regazzini）引用了德 • 菲内蒂（De Finetti）的话：“费希尔丰富而多样的个性表现出一些矛盾。他的共同点是一方面，他缺乏应用意识，另一方面，他对科学研究抱有崇高的理念，这些都使他鄙视真正的客观主义表述的狭隘性，认为这是一种呆板的态度。他通过拒绝贝叶斯-拉普拉斯表述的错误来表明自己坚持客观主义的观点。这里不太好的是他的数学，他在处理个别问题时精通数学，但在处理概念问题时却漫不经心，因此受到了明显而有时严厉的批评。从我们的角度来看，只要我们回到产生这些观察和想法的直觉，并将它们从他认为用来证明它们的论据中解放出来，那么费舍尔的许多观察和想法很可能是正确的。”

图 1 描述了费舍尔统计，它是贝叶斯学派和频率学派之间的一种折衷，但在一个关键方面它并不是折衷：它的易用性。费舍尔的哲学总是以非常实用的方式表达。他似乎很自然地用计算算法来思考，比如最大似然估计、方差分析和排列检验。如果有什么能在 21 世纪取代费舍尔，那必须是一种同样易于在日常实践中应用的方法。

5. 费希尔对当前研究的影响

这次演讲分为三个部分：过去、现在和未来。你们刚才看到的过去部分并没有充分阐述费舍尔的思想，但这里的主题更多的是影响而不是思想，当然，我也承认影响是基于思想的力量。所以现在我要讨论一下费舍尔对当前研究的影响。

以下是最近引起广泛关注的几个（实际上是六个）当前研究课题的例子。这里不声称这些例子是完整的。我试图通过这些例子说明的主要观点是，费舍尔的思想仍然对统计理论的发展产生着强大的影响，这是它们未来相关性的重要迹象。这些例子将逐渐变得更加具有推测性和未来性，并将包括费舍尔未能令人满意地处理的一些发展领域——费舍尔结构中的漏洞——我们可能期望未来的工作在动机上更加倾向于频率论或贝叶斯论。

这些例子也让我有机会谈论统计学家开始关注的新型应用问题，即未来几十年我们将要处理的更大、更混乱、更复杂的数据集。费希尔方法是为了解决 20 世纪 20 年代和 30 年代的问题而发明的。它们是否同样适用于 21 世纪的问题

还不确定——我希望至少能对这个问题有所启发。

5.1 Fisher 信息和引导程序

第一个例子旨在说明 Fisher 的思想如何出现在当前的工作中，但由于计算的进步而难以识别。首先，这里是对 Fisher 信息的简要回顾。假设我们从依赖于单个未知参数 θ 的密度函数 $f_\theta(x)$ 中观察到一个随机样本 $x_1, x_2, \ldots, x_n$，

$$f_\theta(x) \to x_1, x_2, \ldots, x_n$$

任何一个 x 的 Fisher 信息都是 log 密度减去二阶导数的期望值，

$$i_\theta = \mathbf{E}_\theta \left\{ -\frac{\partial^2}{\partial \theta^2} \log f_\theta(x) \right\}$$

并且整个样本的总 Fisher 信息为 ni_θ。

Fisher 证明 MLE 的渐近标准误差与总信息的平方根成反比，

$$\mathrm{se}_\theta(\hat{\theta}) \doteq \frac{1}{\sqrt{ni_\theta}} \tag{4}$$

并且没有其他一致且足够规则的 θ 估计（本质上没有其他渐近、无偏估计量）可以做得更好。

关于统计信息的含义，人们对 i_θ 赋予了大量的哲学解释，但在实践中，Fisher 公式 (4) 最常被用作 MLE 标准误差的简便估计。当然，(4) 本身不能直接使用，因为 i_θ 涉及未知参数 θ。Fisher 的策略看似显而易见，但实际上却是 Fisher 方法论的核心，即在 (4) 中将 MLE $\hat{\theta}$ 代入 θ，从而得到可用的标准误差估计值，

$$\widehat{\mathrm{se}} = \frac{1}{\sqrt{ni_{\hat{\theta}}}}. \tag{5}$$

以下是公式 (5) 的实际应用示例。图 2 显示了一项旨在测试实验性抗病毒药物疗效的小型研究的结果。

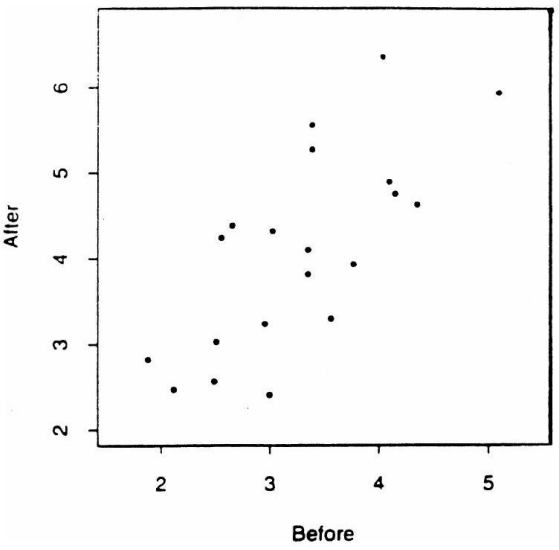

图 2: cd 4 数据；20 名艾滋病患者在服用实验药物之前和之后测量了 cd 4 计数；相关系数 $\hat{\theta} = 0.723$。

总共有 $n = 20$ 名艾滋病患者在服药前后测量了 cd4 计数，得出了数据

$$x_i = (\text{ 之前}_i, \text{ 之后}_i) \text{ 对于} i = 1, 2, \ldots, 20. \ ^{2}$$

皮尔逊样本相关系数为 $\hat{\theta} = 0.723$ 。这个估计有多准确？

如果我们假设数据为双变量正态模型，

$$N_2(\mu, \Sigma) \to x_1, x_2, x_3, \ldots, x_{20}, \qquad (6)$$

表示从具有期望向量 μ 和协方差矩阵 Σ 的二元正态分布中随机抽取 20 对样本的符号，则 $\hat{\theta}$ 是真实相关系数 θ 的 MLE。估计 θ 的 Fisher 信息结果为 $i_\theta = 1/\left(1-\theta^2\right)^2$（在适当考虑了 (6) 中的 “干扰参数” 之后 - 这是我在本次演讲中避免的技术要点之一），因此 (5) 给出了估计的标准误差

$$\widehat{\text{se}} = \frac{\left(1-\hat{\theta}^2\right)}{\sqrt{20}} = 0.107$$

这是针对同一问题的标准误差的自举估计，同样假设双变量正态模型是正确的。在这种情况下，自举样本是从模型 (6) 生成的，但用估计值 $\hat{\mu}$ 和 $\hat{\Sigma}$ 代替未知参数 μ 和 Σ：

$$N(\hat{\mu}, \hat{\Sigma}) \to x_1^*, x_2^*, x_3^*, \ldots, x_{20}^* \to \hat{\theta}^*,$$

其中 $\hat{\theta}^*$ 是自举数据集 $x_1^*, x_2^*, x_3^*, \ldots, x_{20}^*$ 的样本相关系数。

整个过程独立重复了 2,000 次，得到 2,000 个自举相关系数 $\hat{\theta}^*$。图 3 显示了它们的直方图。

2,000 美元 $\hat{\theta}^*$ 值的经验标准差为

$$\widehat{\text{se}}_{\text{boot}} = 0.112$$

这是 $\hat{\theta}$ 2,000 是标准误差所需的 10 倍，但我们稍后将需要全部 2,000 来讨论近似置信区间。

5.2 插件原理

Fisher 信息和 bootstrap 标准误差估计值 0.107 和 0.112 非常接近。这并非偶然。尽管看起来完全不同，但这两种方法进行的计算非常相似。两者都使用 “插件原理” 作为获得答案的关键步骤。

下面是两种方法的插件描述：

- 费舍尔信息 -（i）计算样本相关系数标准误差的（近似）公式，该公式是未知参数 (μ, Σ) 的函数；（ii）将未知参数 $(\mu\ \Sigma)$ 的估计值 $(\hat{\mu}, \hat{\Sigma}\ )$ 代入公式中；
- 引导程序-(i) 将 $(\ \hat{\mu}, \hat{\Sigma})$ 代入生成数据的机制中的未知参数 (μ, Σ)；(ii) 对于代入的机制，通过蒙特卡洛模拟计算样本相关系数的标准误差。

这两种方法的顺序相反，即 “计算然后代入” 与 “代入然后计算”，但这是一个相对较小的技术差异。这两种方法中的关键步骤，也是唯一进行的统计推断，是用估计值 $(\ \hat{\mu}, \hat{\Sigma})$ 代替未知参数 (μ, Σ)，换句话说就是代入原理。费舍尔推断经常使用代入原理，这是费舍尔方法在实践中如此方便的主要原因之一。所有可能的推断问题都可以通过简单地代入未知参数的估计值（通常是最大似然估计值）来回答。

Fisher 信息法涉及的数学知识比 bootstrap 更巧妙，但不得不如此，因为我们比 Fisher 拥有 10^7 的计算优势。一年的综合计算工作量 1925 年所有统计学家的计算时间都比不上现代计算机的一分钟时间。引导法利用这一优势，将 Fisher 的计算扩展到数学变得极其复杂的情况下。Fisher 统计的一个不太吸引人的方面是它

过度依赖一小类简单的参数模型，比如正常模型，考虑到 Fisher 必须使用的机械计算器的局限性，这很容易理解。

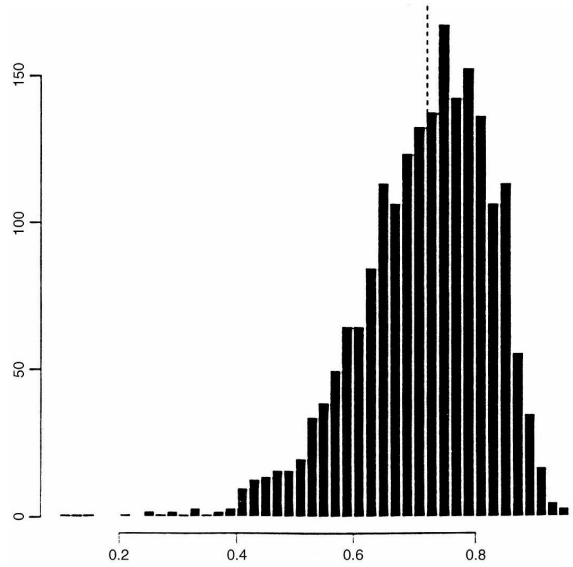

图 3: 2,000 个自举相关系数的直方图；双变量正态抽样模型。

现代计算使我们有机会将 Fisher 的方法扩展到更广泛的模型类别，包括非参数模型（引导法更常见的领域）。我们开始看到许多这样的扩展，例如，将判别分析扩展到 CART，将线性回归扩展到广义加性模型。

6. 标准间隔

我想继续 cd4 的例子，但从标准误差开始讨论置信区间。置信区间的故事说明了如何利用基于计算机的推理以更宏大的方式扩展 Fisher 的思想。

Fisher 使用 MLE 及其估计的标准误差来形成近似置信区间，我喜欢称之为标准区间，因为它们在日常实践中无处不在，

$$\hat{\theta} \pm 1.645\widehat{\mathrm{se}} \tag{7}$$

常数 1.645 为未知参数 θ 提供了大约 90% 覆盖率的区间，区间两端的非覆盖概率为 5%。对于 95% 覆盖率，我们可以使用 1.96 代替 1.645，依此类推，但这里我坚持使用 90%。

标准区间遵循 Fisher 的结果，即 $\hat{\theta}$ 是渐近正态的、无偏的，并且标准误差由样本大小和 Fisher 信息固定，

$$\hat{\theta} \to N\left(\theta, \mathrm{se}^2\right) \tag{8}$$

如 (4) 所示。我们认为 (8) 是 Fisher 理想的 “显而易见” 形式之一。

如果使用决定重要性，那么标准间隔就是费舍尔最重要的发明。它们的流行是由于最优性（或至少是渐近最优性）与计算可处理性。标准间隔为：

- 准确 - 它们的未覆盖概率在间隔的每个端点应该是 0.05，但实际上

$$0.05 + c/\sqrt{n} \tag{9}$$

其中 c 取决于具体情况，因此随着样本大小 n 变大，我们以 $n^{-1/2}$ 的速率接近标称值 0.05；

- 正确 - 基于 Fisher 信息估计的标准误差是任何渐近无偏估计 θ 的最小可能值，因此

区间 (7) 不会浪费任何信息，也不会产生误导性的乐观结果；
- 无论问题多么复杂，自动 $-\hat{\theta}$ 和 $\widehat{\text{se}}$ 都是通过相同的基本算法计算出来的。

尽管有这些优势，应用统计学家知道标准区间在小样本中可能非常不准确。图 4 左侧面板中 cd 4 相关性示例说明了这一点，其中我们看到标准区间端点位于正态理论精确 90% 中心置信区间端点的右侧很远。事实上，我们可以从 bootstrap 直方图（从图 3 复制）中看到，在这种情况下，MLE 的渐近正态性在 $n = 20$ 时尚未成立，因此有充分的理由怀疑标准区间。能够查看包含大量信息的直方图是 Fisher 所没有的奢侈。

Fisher 针对这种特殊情况提出了一个解决方案：将相关系数转换为 $\hat{\phi} = \tanh^{-1}(\hat{\theta})$，即

$$\hat{\phi} = \frac{1}{2}\log\frac{1+\hat{\theta}}{1-\hat{\theta}} \tag{10}$$

在此尺度上应用标准方法，然后将标准区间转换回 θ 尺度。这是 Fisher 的另一个巧妙的简化方法。$\tanh^{-1}$ 变换大大加速了收敛到正态性，正如我们从图 4 右侧面板中 $\hat{\theta}^* = \tanh^{-1}\left(\hat{\theta}^*\right)$ 的 2,000 个值的直方图中看到的那样，并使标准区间更加准确。但是，我们现在已经失去了标准区间的 “自动” 属性。$\tanh^{-1}$ 变换仅适用于正态相关系数，而不适用于大多数其他问题。

标准间隔实际上是大样本近似值 $\hat{\theta} \sim N\left(\theta, \text{se}^2\right)$，即 $\hat{\theta}$ 服从正态分布，对 θ 无偏，且标准误差为常数。更仔细地观察渐近线会发现，这三个假设中的每一个都可能在很大程度上不成立：$\hat{\theta}$ 的抽样分布可能存在偏差；$\hat{\theta}$ 作为 θ 的估计值可能有偏差；其标准误差可能随 θ 而变化。现代计算使纠正这三个错误变得切实可行。我将提到两种方法，第一种使用引导直方图，第二种基于似然法。

事实证明，引导直方图中有足够的信息来纠正所有三个错误

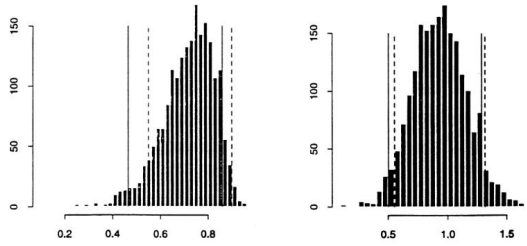

图 4: （左图）$cd4$ 相关系数的精确 90% 置信区间的端点（实线）与标准区间端点（虚线）有很大不同，这由引导直方图的非正态性表明。（右图）Fisher 变换使引导直方图正态化，使标准区间更加准确。标准区间。结果是一个近似置信区间系统，其准确度提高了一个数量级，并且具有非覆盖概率

$$0.05 + c/n$$

与 (9) 相比，实现了所谓的二阶精度。表 1 展示了二阶精度的实际优势。在大多数情况下，我们没有确切的端点作为比较的 “黄金标准”，但二阶精度仍然表明引导间隔的优越性。

bootstrap 方法以及下一节中的基于似然的方法都是变换不变的；也就是说，无论是否进行 $\tanh^{-1}$ 变换，它们都会为相关系数提供相同的区间。从这个意义上讲，它们自动化了 Fisher 的奇妙变换技巧。

表 1: 假设双变量正态性，cd 4 相关系数的精确和近似 90% 置信区间的端点

	Exact	Bootstrap	Standard
0.05	0.464	0.468	0.547
0.95	0.859	0.856	0.899

我喜欢这个例子，因为它展示了基本的费歇尔构造，即标准区间，是如何通过现代计算进行扩展的。通过扩展，我们可以轻松处理非常复杂的概率模型，甚至是非参数模型，以及复杂的统计数据，例如逐步稳健回归中的系数。

此外，扩展不仅限于更广泛的应用。在此过程中，我们在理解近似置信区间的理论基础方面取得了一些进展。其他主题也以同样的方式涌现出来。例如，Fisher 1925 年关于不充分估计量的信息损失的研究已经转化为我们现代的 EM 算法和吉布斯抽样理论。

7. 条件推理、辅助性和魔术公式

表 2 显示了随机实验中非常不良的副作用的发生情况，稍后将对此进行更详细的描述。治疗组产生的不良反应的比例小于对照组，样本对数几率比为

$$\hat{\theta} = \log\left(\frac{1}{15}/\frac{13}{3}\right) = -4.2$$

费舍尔想知道如何对真正的对数几率比 θ 做出适当的推断。这里的问题在于干扰参数。2×2 表的多项式模型有三个自由参数，代表四个单元格概率，这些概率的总和为 1，从某种意义上说，必须消除三个参数中的两个才能得到 θ。为

表 2: 随机实验中不良事件的发生；样本对数比值比 $\hat{\theta} = -4.2$

	Yes	No	
Treatment	1	15	16
Control	13	3	16
	14	18	

此，费舍尔想出了另一种方法，将复杂的情况简化为简单的形式。

Fisher 表明，如果我们以表格的边际为条件，那么给定边际的 $\hat{\theta}$ 的条件密度仅取决于 θ。干扰参数消失。他认为这种条件是 “正确的”，因为边际充当了所谓的近似辅助统计数据。也就是说，它们不携带太多有关 θ 值的直接信息，但它们可以说明 $\hat{\theta}$ 估计 θ 的准确性。后来，Neyman 通过现在所谓的 Neyman 结构，为以边际为条件给出了更具体的频率论依据。

对于表 2 中的数据，给定边际的 $\hat{\theta}$ 条件分布得出 $[-6.3, -2.4]$ 作为 θ 的 90% 置信区间，排除了治疗等于控制的零假设值 $\theta = 0$。然而，即使在这种简单情况下，条件分布也不容易计算，并且在更复杂的情况下会变得难以计算。

在他 1934 年的论文中，他解决了平移族的条件问题，这是 Fisher 在有效估计方面的工作的巅峰之作。假设 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ 是来自柯西分布 (3) 的随机样本，并且我们希望使用 $\mathbf{x}$ 推断 θ，即分布的未知中心点。在这种情况下，有一个真正的辅助统计量 $\mathbf{A}$，即 $\mathbf{x}$ 有序值之间的间距向量。费舍尔再次指出，关于 θ 的正确推论应该基于 $f_\theta(\hat{\theta} \mid \mathbf{A})$，即给定辅助 $\mathbf{A}$ 时 MLE $\hat{\theta}$ 的条件密度，而不是基于无条件密度 $f_\theta(\hat{\theta})$。

Fisher 还提供了一个计算 $f_\theta(\hat{\theta} \mid \mathbf{A})$ 的绝妙技巧。令 $L(\theta)$ 为似然函数：

整个样本的无条件密度，被视为 θ 的函数，其中 $\mathbf{x}$ 固定。然后结果是

$$f_\theta(\hat{\theta} \mid \mathbf{A}) = c\frac{L(\theta)}{L(\hat{\theta})} \tag{11}$$

其中 c 是常数。公式 (11) 允许我们从似然中计算条件密度 $f_\theta(\hat{\theta} \mid \mathbf{A})$，这很容易计算。它还暗示了基于似然的推理（费舍尔的标志）与频率学派方法之间的深刻联系

尽管这个开端很有希望，但在 1934 年之后的几年里，这一承诺未能实现。问题是公式 (11) 只适用于非常特殊的情况，例如不包括 2×2 表格示例。不过，最近人们对基于可能性的条件推理的兴趣又重新燃起。Durbin、Barndorff-Nielsen、Hinkley 等人对 (11) 进行了精彩的推广，适用于具有近似辅助函数的各种问题，即所谓的魔法公式

$$f_\theta(\hat{\theta} \mid \mathbf{A}) = c\frac{L(\theta)}{L(\hat{\theta})}\left\{-\frac{d^2}{d\theta^2}\log L(\theta)\Big|_{\theta=\hat{\theta}}\right\}^{1/2}. \tag{12}$$

在柯西情况下，括号内的因子是常数，将（12）简化回（11）。

基于可能性的条件推理在 Fraser、Cox 和 Reid、McCullagh、Barndorff-Nielson、Pierce、DiCiccio 等人的当前研究中得到了推动。它代表了一项重大努力，旨在完善和扩展 Fisher 的目标，即直接基于可能性的推理系统。

特别是，魔法公式可用于生成比标准区间更准确的近似置信区间，至少是二阶准确。这些区间与引导区间的二阶一致。如果这不是真的，那么它们中的一个或两个就不会是二阶正确的。目前看来，改进标准区间的尝试正在从两个方向汇聚：可能性和引导。

像 (12) 这样的结果具有巨大的潜力。似然推断是费舍尔统计学未实现的重大承诺 - 即一种以同时满足贝叶斯学派和频率学派的方式直接解释似然函数的理论的承诺。即使只是部分实现这一承诺，也会极大地影响 21 世纪统计学的形态。

8. 费舍尔最大的失误

现在，我将小心翼翼地开始进入 21 世纪，讨论一些费舍尔的思想尚未占据主导地位，但在未来发展中可能重要或不重要的话题。我将从基准分布开始，它通常被认为是费舍尔最大的错误。但用亚瑟 • 库斯特勒的话来说，“思想史充满了贫瘠的真理和丰富的错误。” 如果基准推断是一个错误，那么它肯定是一个丰富的错误。

就图 1 中的贝叶斯派和频率派的对比图而言，基准推断是费舍尔最接近贝叶斯派的理论。费舍尔试图在拉普拉斯传统中编纂客观的贝叶斯主义，但不使用拉普拉斯的临时均匀先验分布。我认为费舍尔对基准推断的持续投入有两个主要影响：对奈曼思想的负面反应和对杰弗里斯观点的积极吸引力。

图 5 中的实线是二项式参数 θ 的基准密度，在 10 次试验中观察到 3 次成功，

$$s \sim \text{Binomial}(n, \theta), \quad s = 3 \text{ 和} n = 10$$

图中还显示了一个近似基准密度，稍后我会提到。Fisher 的基准理论最大胆地将实线视为 θ 的真正后验密度，尽管（或者可能是因为）之前没有做出任何假设。

8.1 置信密度

我们也可以把基准分布称为 “置信密度”，因为这是激励基准构造的简单方法。正如我之前所说，Fisher 会讨厌这个名字。

假设对于 0 到 1 之间的每个 α 值，我们都有一个 θ 的上限 100α 置信限度 $\hat{\theta}[\alpha]$，因此根据定义

$$\text{prob}\{\theta < \hat{\theta}[\alpha]\} = \alpha$$

如果我们愿意接受对置信度的经典错误解释，我们可以将其解释为给定数据的 θ 的概率分布，

θ 在区间 $(\hat{\theta}[0.90], \hat{\theta}[0.91])$ 内

概率为 0.01 ，依此类推。

达到连续极限就会得到 “置信密度”，这是奈曼讨厌的一个名字。

置信密度是基准分布，至少在 Fisher 认为置信限度是推断的情况下是如此。

完全正确。图 5 中的基准分布是基于 θ 的通常置信限度的置信密度（考虑到二项分布的离散性质）：$\hat{\theta}[\alpha]$ 是 θ 的值，使得 $S \sim \text{Binomial}(10, \theta)$ 满足由于需要进行临时的连续性校正，费舍尔

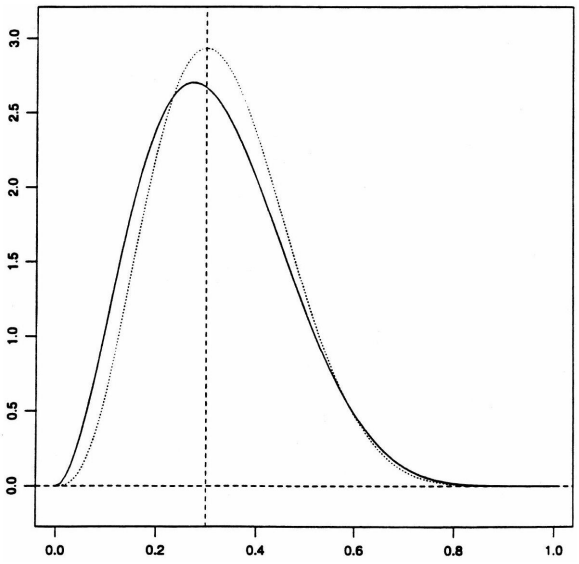

图 5: 二项式参数 θ 的基准密度，在 10 次试验中观察到 3 次成功。虚线是近似值，在复杂情况下很有用。

不愿意将基准论证应用于离散分布，但由此造成的困难更多的是理论上的，而不是实际的。

$$\text{prob}\{S > 3\} + \frac{1}{2}\,\text{prob}\{S = 3\} = \alpha.$$

用置信密度来表述基准思想的优点是，它们可以应用于更广泛的问题。我们可以使用前面提到的近似置信区间（无论是引导区间还是似然区间）来获得近似基准分布，即使在具有大量干扰参数的非常复杂的情况下也是如此。（图 5 中的虚线是基于近似引导区间的置信密度。）并且有实际原因可以说明为什么拥有良好的近似基准分布非常方便，这些原因与我们这个行业 250 年来寻找可靠的客观贝叶斯理论有关。

8.2 客观贝叶斯

我所说的“客观贝叶斯”是指一种贝叶斯理论，其中主观因素从先验分布的选择中被移除；从实践角度来说，这是一种在没有先验信息的情况下应用贝叶斯定理的通用方法。一种被广泛接受的客观贝叶斯理论（基准推断旨在成为这种理论）将具有巨大的理论和实践意义。

我在这里想到的是处理混乱、复杂的问题，我们试图将来自不同来源的信息结合起来——例如进行元分析。贝叶斯方法特别适合解决这类问题。现在，吉布斯采样器和马尔可夫链蒙特卡罗等技术可用于从高维后验分布中整合干扰参数，这一点尤其正确。

当然，问题在于统计学家仍然必须选择一个先验分布才能使用贝叶斯定理。不假思索地使用统一先验并不比拉普拉斯时代好。最近人们投入了大量精力来开发无信息或客观先验分布，这些先验可以安全地消除干扰参数，同时保持对感兴趣的参数的中立性。Kass 和 Wasserman 1996 年的 *JASA* 文章回顾了 Berger、Bernardo 和其他许多人的最新发展，但为高维问题找到真正客观的先验的任务仍然艰巨。

基准分布或置信密度提供了一种巧妙解决这一难题的方法。有一个很好的理由可以证明，置信密度是所有干扰参数以客观方式整合后，感兴趣参数的后验密度。如果这个论点被证明是正确的，那么我们在构建近似置信区间和近似置信密度方面的进展可能会让我们更容易在实际问题中使用贝叶斯思维。

这一切都只是推测，但对于 21 世纪，我可以肯定地预测：统计学家将被要求解决更大、更复杂的问题。我相信，客观贝叶斯方法很有可能被开发用于解决此类问题，而基准推断之类的东西将在这一发展中发挥重要作用。也许 Fisher 最大的失误将在 21 世纪成为轰动一时的事情！

9. 模型选择

模型选择是统计研究的另一个领域，似乎正在取得重要进展，但尚未取得明确突破。这里提出的问题是如何从观察到的数据中选择模型本身，而不仅仅是给定模型的连续参数。F 检验和“ F ”代表 Fisher，有助于完成这项任务，并且无疑是使用最广泛的模型选择技术。然而，即使是相对简单的问题，事情也会很快变得复杂，任何迷失在前向和后向逐步回归程序中的人都可以证明这一点。

事实上，经典的费舍尔估计和测试理论对于模型选择来说是一个良好的开端，但仅此而已。特别是，最大似然估计理论和模型拟合不考虑拟合的自由参数的数量，这就是为什么频率论方法（如 Mallow 的 C_p、赤池信息准则和交叉验证）得以发展。模型选择似乎正在远离其费舍尔根源。

现在统计学家开始看到真正复杂的模型选择问题，有数千甚至数百万个数据点和数百个候选模型。机器学习这一蓬勃发展的领域已经发展起来，用于处理此类问题，但其方式与统计理论还没有很好的联系。

表 3 取自 Gail Gong 1982 年的论文，显示了模型选择问题的部分数据，该问题按今天的标准来看只是中等复杂度，但从战前的角度来看却非常困难。对 155 名慢性肝炎患者的“训练集”

进行了 19 个诊断预测变量的测量。结果变量 y 是患者是否死于肝功能衰竭（122 人存活，33 人死亡），该研究的目标是根据诊断变量制定 y 的预测规则。

为了预测结果，我们分三步建立了逻辑回归模型：

- 对 19 个预测因子中的每一个都进行了单独的逻辑回归，结果显示 13 个预测因子在 0.05 水平上具有显著性。
- 前向逐步逻辑回归程序（仅包括 13 个预测因子均未缺失的患者）在重要性水平 0.10 下保留了 13 个预测因子中的 5 个。
- 第二个前向逐步逻辑回归程序，包括 5 个预测因子均未缺失的患者，在显著性水平 0.05 下保留了 5 个预测因子中的 4 个。

最后四个变量，

(13) 腹水, (15) 胆红素,
(7) 不适, (20) 组织学,

被视为 “重要预测因素”。基于这些因素的逻辑回归将 155 名患者中的 16% 错误分类，交叉验证表明真实错误率约为 20%。

一个关键问题涉及所选模型的有效性。从医学角度看，我们是否应该非常认真地对待这四个 “重要预测因素”？引导式答案似乎是 “可能不会”，尽管医学研究人员这样做是理所当然的，因为他们在选择过程中使用了大量的统计机制。

Gail Gong 对 155 名患者进行了重新采样，将每位患者的全部 19 个预测因子和响应记录作为一个单位。对于每个包含 155 个重新采样记录

表 3: 对 155 名慢性肝炎患者的 19 个诊断变量进行了测量；数据显示的是最后 11 名患者的数据；结果 y 为 0 或 1，表示患者存活或死亡；负数表示数据缺失

y	Constant 1	Age 2	Sex 3	Steroid 4	Antiviral 5	Fatigue 6	Malaise 7	Anorexia 8	Liver Big 9	Liver Firm 10	Spleen Palp 11	Spiders 12	Ascites 13	Varices 14	Bilirubin 15	Alk Phos 16	SGOT 17	Albumin 18	Protein 19	Histology 20	#
1	1	45	1	2	2	1	1	1	2	2	2	1	1	2	1.90	−1	114	2.4	−1	−3	145
0	1	31	1	1	2	1	2	2	2	2	2	2	2	2	1.20	75	193	4.2	54	2	146
1	1	41	1	2	2	1	2	2	2	1	1	1	2	1	4.20	65	120	3.4	−1	−3	147
1	1	70	1	1	2	1	1	1	−3	−3	−3	−3	−3	−3	1.70	109	528	2.8	35	2	148
0	1	20	1	1	2	2	2	2	2	−3	2	2	2	2	0.90	89	152	4.0	−1	2	149
0	1	36	1	2	2	2	2	2	2	2	2	2	2	2	0.60	120	30	4.0	−1	2	150
1	1	46	1	2	2	1	1	1	2	2	2	1	1	1	7.60	−1	242	3.3	50	−3	151
0	1	44	1	2	2	1	2	2	2	1	2	2	2	2	0.90	126	142	4.3	−1	2	152
0	1	61	1	1	2	1	1	2	1	1	2	1	2	2	0.80	95	20	4.1	−1	2	153
0	1	53	2	1	2	1	2	2	2	2	1	1	2	1	1.50	84	19	4.1	48	−3	154
1	1	43	1	2	2	1	2	2	2	2	1	1	1	2	1.20	100	19	3.1	42	2	155

的 bootstrap 数据集，她重新运行了三阶段逻辑回归模型，得到了一组 “重要预测因子”。这进行了 500 次。图 6 显示了最后 25 个 bootstrap 数据集的重要预测因子。其中第一个是 $(13, 7, 20, 15)$，除了顺序之外，与原始数据中的集合 $(13, 15, 7, 20)$ 一致。这在 499 个 bootstrap 案例中的任何其他案例中都没有发生。在所有 500 次引导重复中，只有变量 20（组织学）出现了 295 次，在超过一半的时间内是 “重要的”。这些结果无疑削弱了人们对预测变量（13,15,7,20）的因果性质的信心。

或者说他们真的会这么做？看来我们应该能够使用引导结果来定量评估各种预测因子的有效性。也许它们还可以帮助选择更好的预测模型。这些天来，人们一直在问这样的问题，但迄今为止，答案更有趣，而不是结论性的。

我不清楚 Fisher 方法是否会在模型选择理论的进一步发展中发挥重要作用。图 6 使模型选择看起来像离散估计的练习，而 Fisher 的 MLE 理论始终针对连续情况。交叉验证等直接频率论方法目前似乎更有前景，贝叶斯模型选择也有一些最新进展，但事实上，我们迄今为止的最大努力不足以解决肝炎数据等问题。我们非常需要巧妙的 Fisher 技巧，将复杂的模型选择问题简化为简单明显的问题。

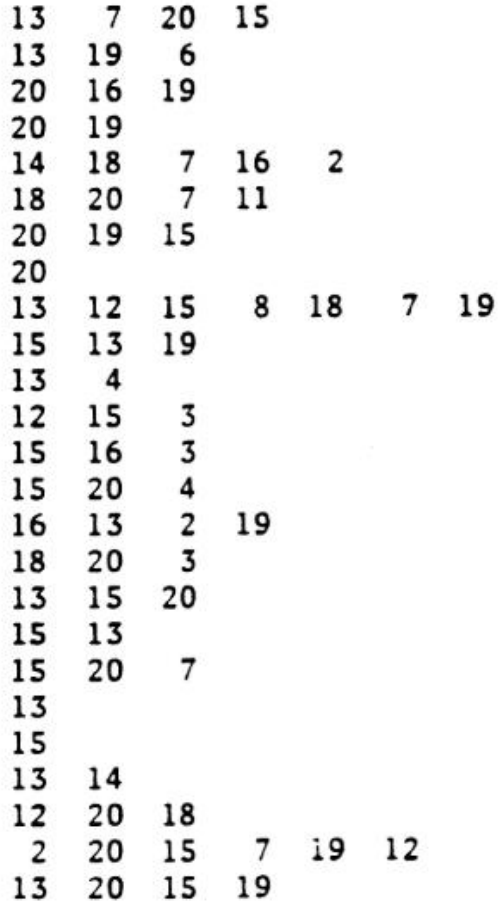

图 6: 三步逻辑回归模型选择程序的 500 个引导重复中的最后 25 个中选择的 “重要预测因子” 集；最初的选择是 $(13, 15, 7, 20)$ 。

10. 经验贝叶斯方法

作为最后一个例子，我想谈谈经验贝叶斯方法。经验贝叶斯在我看来，这似乎是未来的潮流，但 25 年前看起来就是这样，尽管这是一个具有巨大潜在重要性的领域，但潮流仍未席卷而来。这

不是费希尔 (Fisher) 投入太多的话题。

表 4 显示了经验贝叶斯情况的数据：在 41 个城市进行了独立临床试验，比较了两种胃溃疡手术技术（一种新疗法和一种旧疗法）的复发性出血（不良副作用）的发生率。每项试验都得出了复发性出血的真实 log 几率比的估计值，治疗组与对照组，

$$\theta_i = \log \text{ 城市中的几率比} i, \quad i = 1, 2, \ldots, 41.$$

例如，在城市 8 中，我们有表 2 中所示的估计值，

$$\hat{\theta} = \log\left(\frac{1}{15}/\frac{13}{3}\right) = -4.2$$

这表明新手术在减少复发性出血方面非常有效，至少在城市 8

图 7 显示了 41 个城市中 10 个城市的 θ_i 的似然值。这些是条件似然值，使用 Fisher 的条件化边际技巧来消除每个城市的干扰参数。很明显，对数几率比 θ_i 并不完全相同。例如，城市 8 和 13 的似然值几乎没有重叠。另一方面，θ_i 值并没有太大差异，41 个城市中的大多数似然函数集中在范围（-6,3）上。（这是我在讨论基准推断、置信密度和客观贝叶斯方法时担心的那种复杂推理情况。）

请注意，L_8（θ_8 的似然值）位于大多数其他曲线的左侧。如果我们能看到所有 41 条曲线而不是仅 10 条曲线，情况仍然如此。换句话说，θ_8 似乎比其他大多数城市的对数几率比更负。

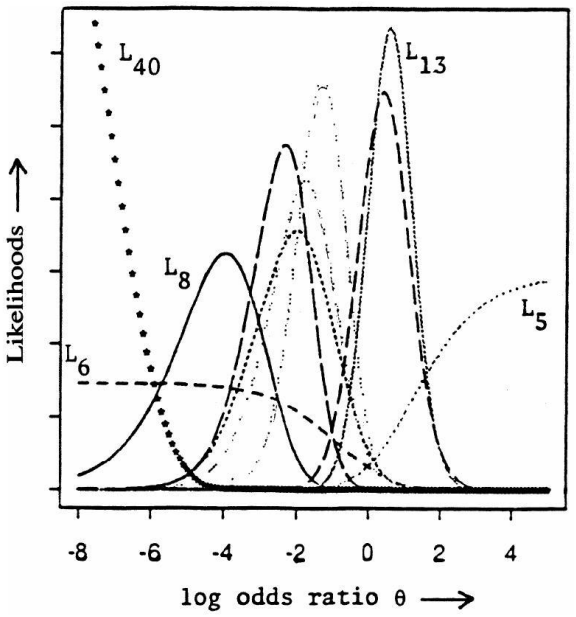

图 7: 表 $4 中的 41 个实验中 10 个的 θ_i 的个体似然函数；L_{8}$（城市 8 的对数优势比的似然）位于大多数其他城市的左侧。

表 4: 溃疡数据：41 项独立实验，涉及溃疡手术后复发性出血的次数；治疗（一种新的手术技术）中 $(a, b) = (\#\text{ 出血}; \#\text{ 非出血})$；$(c, d)$ 与对照组相同，后者是较老的手术；$\hat{\theta}$ 是样本对数优势比，其中估计标准差 $\widehat{\text{SD}}$; 星号表示图 7 中所示的情况

Experiment	a	b	c	d	$\hat{\theta}$	$\widehat{\text{SD}}$	Experiment	a	b	c	d	$\hat{\theta}$	$\widehat{\text{SD}}$
1*	7	8	11	2	−1.84	0.86	21	6	34	13	8	−2.22	0.61
2	8	11	8	8	−0.32	0.66	22	4	14	5	34	66	0.71
3*	5	29	4	35	0.41	0.68	23	14	54	13	61	20	0.42
4	7	29	4	27	0.49	0.65	24	6	15	8	13	−43	0.64
5*	3	9	0	12	inf	1.57	25	0	6	6	0	−inf	2.08
6*	4	3	4	0	−inf	1.65	26	1	9	5	10	−1.50	1.02
7*	4	13	13	11	−1.35	0.68	27	5	12	5	10	−0.18	0.73
8*	1	15	13	3	−4.17	1.04	28	0	10	12	2	−inf	1.60
9	3	11	7	15	−0.54	0.76	29	0	22	8	16	−inf	1.49
10*	2	36	12	20	−2.38	0.75	30	2	16	10	11	−1.98	0.80
11	6	6	8	0	−inf	1.56	31	1	14	7	6	−2.79	1.01
12*	2	5	7	2	−2.17	1.06	32	8	16	15	12	−0.92	0.57
13*	9	12	7	17	0.60	0.61	33	6	6	7	2	−1.25	0.92
14	7	14	5	20	0.69	0.66	34	0	20	5	18	−inf	1.51
15	3	22	11	21	−1.35	0.68	35	4	13	2	14	0.77	0.87
16	4	7	6	4	−0.97	0.86	36	10	30	12	8	−1.50	0.57
17	2	8	8	2	−2.77	1.02	37	3	13	2	14	0.48	0.91
18	1	30	4	23	−1.65	0.98	38	4	30	5	14	−0.99	0.71
19	4	24	15	16	−1.73	0.62	39	7	31	15	22	−1.11	0.52
20	7	36	16	27	−1.11	0.51	40*	0	34	34	0	−inf	2.01
							41	0	9	0	16	NA	2.04

对于 θ_8 来说，一个好的估计值或置信区间是多少？回答这个问题取决于其他城市的结果对我们关于城市 8 的思考有多大的影响。这就是经验贝叶斯理论的作用所在，它为我们提供了一个系统框架，用于将城市 8 的实验中关于 θ_8 的直接信息与其他 40 个城市实验中的间接信息结合起来。

仅基于其自身实验的数据 $(1, 15, 13, 3)$，θ_8 的普通 90% 置信区间为

$$\theta_8 \in [-6.3, -2.4]. \tag{13}$$

经验贝叶斯方法给出了截然不同的结果。经验贝叶斯分析使用其他 40 个城市的数据来估计对数几率比的先验密度。使用贝叶斯定理，可以将该先验密度与城市 8 的似然值 L_8 相结合，得到 θ_8 的中心 90% 后验区间

$$\theta_8 \in [-5.1, -1.8]. \tag{14}$$

大多数城市的负面趋势结果都比城市 8 要小，这一事实在实证贝叶斯分析中发挥了重要作用。从其他 40 个城市估计的贝叶斯先验表明，θ_8 不太可能像其自身数据所显示的那么负面。

经验贝叶斯分析意味着，其他 40 个城市的数据中有很多信息可用于估计 θ_8，事实上，与城市 8 自己的数据中的信息差不多。这种 “其他” 信息没有明确的费舍尔解释。整个经验贝叶斯分析都是贝叶斯主义的，就好像我们从 θ_8 的真正信息丰富的先验开始，但它仍然声称具有频率主义的客观性。

也许我们正濒临贝叶斯方法和频率学派之间的新折衷，这种折衷与费舍尔的提议有着根本的不同。如果是这样，那么 21 世纪可能看起来就不那么像费舍尔了，至少对于溃疡数据等具有平行结构的问题而言是如此。目前，这样的问题并不多。如果统计学界对分析经验贝叶斯问题更有信心，这种情况可能会很快改变。在费舍尔提供有效的方法来处理因子设计问题之前，并没有太多的因子设计问题。科学家倾向于给我们带来我们可以解决的问题。目前对元分析和分层模型的关注无疑表明人们对经验贝叶斯这类情况的兴趣日益浓厚。

11. 统计三角

现代统计理论的发展是贝叶斯、频率论和费舍尔三方观点的较量。我试图用我的例子来说明这三种哲学对当前几个热门话题的影响：标准误差估计；近似置信区间；条件推理；客观贝叶斯理论和基准推理；模型选择；经验贝叶斯技术。

图 8 统计三角形更简洁地说明了这一点。它使用重心坐标来表示贝叶斯、频率论和费舍尔思想对各种活跃研究领域的影响。三角形的费舍尔极位于贝叶斯极和频率论极之间，如图 1 所示，但在这里我为费舍尔哲学分配了一个自己的维度，以考虑其独特的操作特征：简化为 “明显” 类型；插入原则；强调推理正确性；直接解释可能性；以及使用自动计算算法。

当然，即使接受作者的偏见，这样的图景也只能大致准确，但许多位置很难争论。我毫不费力地将条件推理和部分可能性放在费舍尔极点附近，

将稳健性放在频率极点附近，将多重插补放在贝叶斯极点附近。经验贝叶斯显然是贝叶斯和频率思想的混合体。引导方法将插入原理的便利性与频率主义者对精确操作特性的强烈渴望结合起来，特别是对于近似置信区间，而折刀法的发展则更加纯粹地体现了频率主义。

图 8 中的其他一些位置更成问题。Fisher 提供了 EM 算法背后的原始思想，事实上，最大似然估计的自洽性（当缺失数据由统计学家填写时）是经典的 Fisher 正确性论证。另一方面，EM 的现代发展具有强大的贝叶斯成分，在吉布斯抽样的相关主题中可以更清楚地看到。同样，Fisher 组合独立 p 值的方法是元分析的早期形式，但该主题最近的发展一直是强烈的频率论。

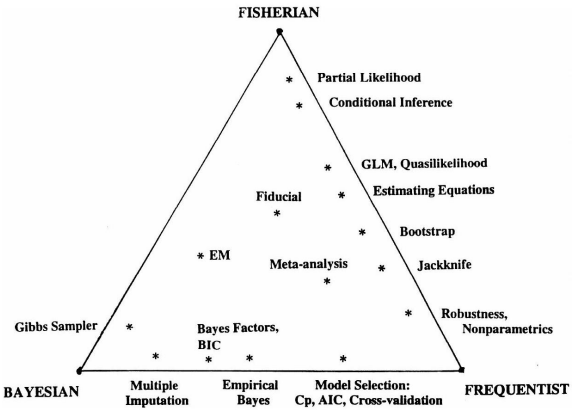

图 8: 现代统计研究的重心图，显示了贝叶斯、频率学派和费舍尔哲学对当前感兴趣的各种主题的相对影响。

这里的问题是，费舍尔并不总是费舍尔主义者，因此很容易将血统与发展相混淆。

最困难和最尴尬的情况涉及我一直称之为“客观贝叶斯”的方法，其中包括基准推断。频率论的一个定义是希望针对每个可能的先验分布都做得很好，或者至少不要做得很差。贝叶斯精神，正如 Savage 和 de Finetti 所体现的那样，就是针对一个先验分布（大概是正确的）做得非常好。

在这两个极端之间，存在各种客观的贝叶斯折衷方案。韦尔奇和皮尔斯在频率论的极端附近工作，他们展示了如何计算先验，其后验可信区间与标准置信区间紧密吻合。杰弗里斯的工作导致了贝叶斯模型选择的现代蓬勃发展，但其频率论色彩较淡。在双变量正态情况下，杰弗里斯会建议使用相同的先验分布来估计相关系数或期望比，而韦尔奇-皮尔斯理论会使用两个不同的先验来分别匹配每个频率论解决方案。

尽管如此，杰弗里斯的贝叶斯主义具有不可否认的客观主义色彩。埃里希 • 莱曼（个人通信）曾这样说过：“如果将两个贝叶斯概念 [Savage-de Finetti 和 Jeffreys] 分开，并只将主观版本放在贝叶斯角落，在我看来，会发生一些有趣的事情：杰弗里斯的概念向右移动，最终更接近频率角落，而不是贝叶斯角落。例如，你将贝叶斯的乐观和冒险与频率主义者的悲观和安全作斗争进行了对比。在这两个尺度上，杰弗里斯更接近频率主义者的一端。事实上，无信息先验的概念在哲学上接近沃尔德最不利的分布，而且两者经常重合。”

图 8 稍微遵循了 Lehmann 的建议，其中贝叶斯模型选择 (BIC) 点（Jeffreys 工作的直接遗产）已稍微向频率论极点移动。但是，我将基准推断（Fisher 形式的客观贝叶斯主义）定位在三角形的中心附近。目前该领域的工作并不多，但来自

三个方向的需求都很大。

我举例的目的，也是这次演讲的重点，是要表明费舍尔统计并非一种消亡的语言，它继续激发新的研究。我认为图 8 中这一点很明显，即使考虑到它的不准确性。但费舍尔的语言并不是城里唯一的语言，它甚至不是我们研究期刊的主导语言。这个奖项必须颁给一种相当随意的频率主义，如今通常不像纯决策理论那样尖锐。我们可能会问 20 或 30 年后图 8 会是什么样子，三角形的费希尔极附近是否有许多活跃的研究兴趣点。

12. 21 世纪的 R. A. 费希尔

大多数关于未来的讨论实际上都是关于现在的，这次当然也不例外。但在这次讨论结束、20 世纪即将结束之际，我们可以谨慎地展望未来，至少对费舍尔统计的未来进行一点推测。

当然，Fisher 的基本发现，如充分性、Fisher 信息、MLE 的渐近效率、实验设计和随机化推理，都不会消失。但它们可能会变得不那么显眼。目前，我们几乎完全按照 Fisher 创造的方式使用这些想法，但现代计算设备可能会改变这一点。

例如，在涉及大量干扰参数的某些情况下，最大似然估计可能会严重偏差（如奈曼-斯科特悖论）。偏差较小的计算机修改版 MLE 可能成为应用问题中默认的首选估计量。方差分量的 REML 估计提供了一个当前的例子。同样，随着高性能计算机的普及，统计学家可能会自动使用我之前提到的某种更准确的置信区间，而不是标准区间。

类似这样的变化会掩盖费舍尔的影响，但不会真正削弱它。不过，有几个很好的理由可以让我们期待统计领域发生更剧烈的变化，其中第一个是我们的计算设备每十年都会有奇迹般的改进，改进幅度是数量级的。设备是科学的命运，统计学也不例外。其次，统计学家被要求解决更大、更难、更复杂的问题，比如模式识别、DNA 筛选、神经网络、成像和机器学习。统计领域的新问题总是会引发新的解决方案，但这一次的解决方案可能必须非常激进。

几乎从定义上来说，预测根本性的变化都是困难的，但我想以一些关于统计学未来的推测性可能性作为结束，这些推测性可能性可能会、也可能不会大大减少费希尔的统计意义。

12.1 贝叶斯世界

1974 年，丹尼斯 • 林德利 (Dennis Lindley) 预测 21 世纪将是贝叶斯世纪。（我注意到他最近的统计科学访谈现在预测 2020 年。）他可能是对的。贝叶斯方法对于像刚才提到的那些复杂问题很有吸引力，但除非科学界改变思维方式，否则我无法想象主观贝叶斯方法会占据主导地位。我所说的客观贝叶斯，即使用中性或无信息的先验，似乎更有希望，而且这些天肯定很流行。

成功的客观贝叶斯理论必须在熟悉的情况下提供良好的频率学特性，例如，对于替代置信区间的任何事物，都有合理的覆盖概率。除了对多重比较等复杂问题的更直接分析外，这样的贝叶斯世界似乎与当前情况没有太大不同。我们可以想象 2020 年的统计学家弯腰驼背地坐在超

级计算机终端上，试图让 Proc Prior 成功运行，我们只能祝愿未来的同事 “好运”。

12.2 非参数

作为我们费舍尔传统的一部分，我们倾向于过度使用简单的参数模型，就像正常情况一样。在非参数世界中，参数模型是最后的手段，而不是首选，这将有利于三角形图像中的频率顶点。

12.3 新的综合

战后，尤其是最近几十年，方法论的进步比统计推断理论中基本新思想的发展更为显著。这并不意味着这种发展永远结束了。费舍尔的工作在 20 世纪 20 年代突然出现，也许我们的领域即将迎来另一道闪电。

我们很容易想象，费舍尔、奈曼等人很幸运，因为他们生活在一个好主意还没有被从树上摘下来的时代。事实上，我们正生活在统计学的黄金时代——计算变得快速而简单。从这个意义上说，我们早就应该有一个新的统计范式来巩固战后时期的方法论成果。用琼 • 费舍尔 • 博克斯的比喻来说，废墟又在堆积起来，我们非常需要一个新的费舍尔来让我们的世界井然有序。

我的实际猜测是老费舍尔将有一个非常好的 21 世纪。应用统计学的世界似乎需要在贝叶斯和频率论思想之间进行有效的妥协，而目前还看不到任何可以替代费舍尔综合理论的东西。此外，费舍尔的理论非常适合计算机时代。Fisher 似乎很自然地以算法的方式思考。最大似然估计、标准区间、方差分析表、置换检验都以算法的方式表达，并且很容易通过现代计算进行扩展。

最后我要说的是，费舍尔是一流的天才，他完全有资格成为 20 世纪最重要的应用数学家。他的工作具有大胆的数学综合与极致实用相结合的独特品质。他的作品在我们这个领域留下了深刻的印记，并且没有褪色的迹象。这是一位伟大思想家的印记，统计学——以及一般科学——都深受他的影响。

参考文献

Section 1

SavaGe, L. J. H. (1976). On rereading R. A. Fisher (with discussion). Ann. Statist. 4 441-500. (Savage says that Fisher's work greatly influenced his seminal book on subjective Bayesianism. Fisher's great ideas are examined lovingly here, but not uncritically.)

Yates, F. and Mather K. (1971). Ronald Aylmer Fisher. In Collected Papers of R. A. Fisher (K. Mather, ed.) 1 23-52. Univ. Adelaide Press. (Reprinted from a 1963 Royal Statistical Society memoir. Gives a nontechnical assessment of Fisher's ideas, personality and attitudes toward science.)

Box, J. F. (1978). The Life of a Scientist. Wiley, New York. (This is both a personal and an intellectual biography by Fisher's daughter, a scientist in her own right and also an historian of science, containing some unforgettable vignettes

of precocious mathematical genius mixed with a difficulty in ordinary human interaction. The sparrow quote in Section 4 is put in context on page 130.)

Section 2

Fisher, R. A. (1925). Theory of statistical estimation. Proc. Cambridge Philos. Soc. 22 200-225. (Reprinted in the Mather collection, and also in the 1950 Wiley Fisher collection Contributions to Mathematical Statistics. This is my choice for the most important single paper in statistical theory. A competitor might be Fisher's 1922 Philosophical Society paper, but as Fisher himself points out in the Wiley collection, the 1925 paper is more compact and businesslike than was possible in 1922, and more sophisticated as well.)

Efron B. (1995). The statistical century. Royal Statistical Society News 22 (5) 1-2. (This is mostly about the postwar boom in statistical methodology and uses a different statistical triangle than Figure 8.)

Section 3

FISHER, R. A. (1934). Two new properties of mathematical likelihood. Proc. Roy. Soc. Ser. A 144 285-307. (Concerns two situations when fully efficient estimation is possible in finite samples: one-parameter exponential families, where the MLE is a sufficient statistic, and location-scale families, where there are exhaustive ancillary statistics. Reprinted in the Mather and the Wiley collections.)

Section 4

Efron, B. (1978). Controversies in the foundations of statistics. Amer. Math. Monthly 85 231-246. (The Bayes-Frequentist-Fisherian argument in terms of what kinds of averages should the statistician take. Includes Fisher's famous circle example of ancillarity.)

Efron, B. (1982). Maximum likelihood and decision theory. Ann. Statist. 10 240-356. (Examines five questions concerning maximum likelihood estimation: What kind of theory is it? How is it used in practice? How does it look from a frequentistic decision-theory point of view? What are its principal virtues and defects? What improvements have been suggested by decision theory?)

Cifarelli, D. and Regazzini, E. (1996). De Finetti's contribution to probability and statistics. Statist. Sci. 11 253-282. [The second half of the quote in my Section 4, their Section 3.2.2, goes on to criticize the Neyman-Pearson school. De Finetti is less kind to Fisher in the discussion following Savage's (1976) article.]

Sections 5 and 6

DiCiccio, T. and Efron, B. (1996). Bootstrap confidence intervals (with discussion). Statist. Sci. 11 189-228. (Presents and discusses the cd 4 data of Figure 2. The bootstrap confidence limits in Table 1 were obtained by the BC_a method.)

Section 7

REID, N. (1995). The roles of conditioning in inference. Statist. Sci. 10 138-157. [This is a survey of the p^* formula, what I called the magic formula following Ghosh's terminology, and many other topics in conditional inference; see also the discussion (following the companion article) on pages 173-199, in particular McCullagh's commentary. Gives an extensive bibliography.]

Efron, B. and Hinkley, D. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. Biometrika 65 457-487. (Concerns ancillarity, approximate ancillarity and the assessment of accuracy for a MLE.)

Section 8

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. Biometrika $803 - 26$. (Shows how approximate confidence intervals can be used to get good approximate confidence densities, even in complicated problems with a great many nuisance parameters.)

Section 9

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. Amer. Statist. 37 36-48. (The chronic hepatitis example is discussed in Section 10 of this bootstrap-jackknife survey article.)

O'HagAN, A. (1995). Fractional Bayes factors for model comparison (with discussion). J. Roy. Statist. Soc. Ser. B 57 99-138. (This paper and the ensuing discussion, occasionally rather heated, give a nice sense of Bayesian model selection in the Jeffreys tradition.)

KASs, R. and RAFtERY, A. (1995). Bayes factors. J. Amer. Statist. Soc. 90 773-795. (This review of Bayesian model selection features five specific applications and an enormous bibliography.)

Section 10

ErRon, B. (1996). Empirical Bayes methods for combining likelihoods. J. Amer. Statist. Assoc. 91 538-565

Section 11

KASS, R. and WASSERmAN, L. (1996). The selection of prior distributions by formal rules.

J. Amer. Statist. Assoc. 91 1343-1370. (Begins “Subjectivism has become the dominant philosophical tradition for Bayesian inference. Yet in prac- tice, most Bayesian analyses are performed with so-called noninformative priors. …”)

Section 12

Lindley, D. V. (1974). The future of statistics-a Bayesian 21st century. In Proceedings of the Conference on Directions for Mathematical Statistics. Univ. College, London.

Smith, A. (1995). A conversation with Dennis Lindley. Statist. Sci. 10 305-319. (This is a nice view of Bayesians and Bayesianism. The 2020 prediction is attributed to de Finetti.)

评论

D. R. Cox[1]

我非常喜欢 Efron 教授对 R. A. Fisher 的贡献及其现实意义所作的雄辩而敏锐的评价。我确信，埃夫隆教授对费雪思想的高度重视是正确的。

正如埃夫隆教授所强调的那样，费希尔的思想非常广泛，不可能在一篇论文中全部涵盖。以下提纲挈领的说明是对该论文的补充而非异议。

[1]D. 考克斯（D. R. Cox）是英国牛津纳菲尔德学院（Nuffield College, Oxford OX1 1NF）的荣誉研究员（电子邮箱：david.cox@nuf.ox.ac.uk）。

(1) 费雪强调需要对不同类型的推论问题采取不同的攻击模式。

(2) 虽然他的正式观点涉及的是完全参数问题，但他根据设计中使用的程序给出了“精确”的随机化检验。对他来说，这种检验的地位并不完全清楚。他是将其视为对正态性工作假设感到胆怯的胆小者的定心丸，还是将其视为基于正态理论的结果往往是一种方便近似的首选分析方法？耶茨强烈反对第二种解释。更重要的一点可能是，费雪认识到随机化表明了适当的方差分析，即适当的误差估计。这取代了为每种设计建立新的线性模型的特殊假设。

(3) 要理解费雪方法和奈曼-皮尔逊方法之间的某些区别，关键在于费雪对“唯一”事件的概率 p 的含义的特殊概念，例如费雪（1956 年，第 31-36 页）中所阐述的概念。这包括两个方面，一是个体属于一个集合或种群，其比例为 p，事件在该集合或种群中成立；二是不可能认识到个体属于比例不同的子种群。费雪认为（在我看来是正确的），这使得在没有其他信息可用的情况下，可以根据随机抽样对未知参数附加概率声明，而无需调用先验分布。问题在于，这种分布不能用普通的概率法来操纵或组合。

(4) 通过比较费雪在《实验设计》（Design of Experiments）（费雪，1935 年）一书开头的论战言论与费雪（1956 年）中更为谨慎的评论，可以追溯到费雪贝叶斯推断思想的发展历程。

(5) 当然，费舍尔是一位非常厉害的数学家，尤其是在分布和组合计算方面。Mahalanobis

(1938) 写道：“对严谨论证的明确表述使人感兴趣：

> 他写道：“他对严谨论证的明确表述感兴趣，但前提是必须明确证明这种严谨性。对他来说，机械地演练严谨论证的技巧是令人憎恶的，这一方面是因为迂腐，另一方面也是对积极运用思维的一种抑制。他认为更重要的是积极思考，即使偶尔出错也在所不惜，因为机敏的智慧很快就能从中恢复过来，而不是借助设计最完美的机械拐杖，以蜗牛般的速度沿着众所周知的路径完全安全地前进。

这是对费舍尔在剑桥大学读本科时的态度的评论；人们很容易认为马哈拉诺比斯是在用费舍尔自己的话。

(6) 在某些方面，Fisher 最伟大的方法论贡献，包括判别分析在内的方差分析，以及关于实验设计的思想，似乎与他对一般统计理论的思想没有多少直接关系。当然，后来人们会发现其中存在一些联系，例如与充分性和转换模型的联系，以及在随机化的情况下与条件和辅助统计之间的联系，但这些联系还不够充分。然而，Fisher 对分布理论的掌握显然与此有关，也许最引人注目的是他对多元回归分布理论与线性判别分析分布理论之间联系的研究。

(7) 在正常的科学发展过程中，重要的概念会被简化并融入到该学科的一般精神中，除了思想史学家之外，参考原始资料就变得没有必要了。费舍尔的思想范围之广和精妙之处就在于，阅读上述两本书的至少一部分以及埃夫隆教授在参考文献中提到的论文仍然大有裨益。

（8）我同意 Efron 教授的观点，即当前的一个关键问题是综合关于修改的似然函数、经验贝叶斯方法和一些参考先验概念的思想。

(9) 我喜欢埃夫隆教授的三角形，尽管总的来说我更喜欢正方形，一个轴代表数学公式，另一个轴代表概念目标。例如，费舍尔和杰弗里斯的目标几乎相同，尽管他们的数学当然不同。

（10）最后，在思考费舍尔的贡献时，人们不能忘记，他是一位杰出的遗传学家，也是一位杰出的统计学家。

评论

Rob Kass [2]

非常好，非常发人深省——当我读到埃夫隆版本的费舍尔对我们学科的深远影响时，我不禁想，如果统计学主要基于贝叶斯逻辑而不是主要基于费舍尔频率论逻辑，它是否能够在如此多的学科中取得如此惊人的成功。如果没有历史反事实，这个问题可能会这样表述：从贝叶斯的角度来看，频率论逻辑何时是必要的？如果有的话。

我认为频率论推理对于我们事业的成功至关重要的有三个方面。首先，我们需要拟合优度评估，或者 Dempster 在他的 Fisher 反思中称之为“后验推理”（参见 Gelman、Meng 和 Stern，

[2]Rob Kass 是卡内基梅隆大学统计系教授兼系主任，宾夕法尼亚州匹兹堡 15213（电子邮件：kass@stat.cmu.edu）。

1996 年，以及对此的讨论）。其次，尽管已经制定了许多定义非信息性或参考性先验的原则，但似乎重复采样下的良好行为应该在某种程度上发挥作用。正如 Cox 在 1994 年《统计科学》采访中所暗示的那样，我们可能还没有认识到如何做到这一点，我觉得 Efron 也持这种观点。但第三个也许是最重要的方面，即使是我们贝叶斯主义者目前也认为频率论方法很有用，因为它们提供了非常理想的捷径。这与 Efron 在第 4 节末尾的观点有关。对于现在相对简单的情况，例如方差分析，Fisher 频率论方法当然易于使用；目前尚不清楚标准化贝叶斯方法是否可以完全取代它们。然而，同样重要的是更为复杂的数据分析方法，例如现代非参数回归，它们在易用性方面比贝叶斯方法具有很大的相对优势。简而言之，尽管我强烈偏爱贝叶斯思维，但基于我们目前对推理的理解，我无法想象下一个世纪或任何其他世纪会完全是贝叶斯的。长期以来，我一直认为贝叶斯与频率论的对比，虽然严格来说是一个逻辑问题，但更有用的是，从隐喻的角度看，是一个语言问题——也就是说，它们是两种用来应对不确定性的替代语言，流利的讲者尽管某些短语没有很好的翻译，但一方能够理解另一方流利使用者能够理解的任何现象。

我想我们都同意费舍尔的伟大。我想说费舍尔之于统计学就像牛顿之于物理学。继续这个类比，埃夫隆认为我们需要一个统计上的爱因斯坦。但真正的问题是是否有可能获得一个实现基准推断目标的新框架。统计推断的情况对于单参数问题来说非常平静和令人信服：参考贝叶斯和费舍尔路径通过魔法公式收敛到二阶。然而，当我们进入多参数世界时，不仅贝叶斯和频率论范式的调和的希望渺茫，而且频率论或贝叶斯框架中任何令人满意的统一方法的希望也变得渺茫，我们必须怀疑这个世界是否只是令人沮丧的混乱。事实上，我在 1996 年与拉里·沃瑟曼合写的 JASA 评论论文中发出的一些警告暗示（正如纯粹的主观主义者很快就会争辩的那样）可能没有办法解决根本的困难。

显然，统计问题正变得越来越复杂。我可能错误地认为（从第 8 节末尾的评论中）Efron 将此与他对当前形势的不安以及对新范式的需求联系起来——也许由我们的爱因斯坦救世主提供。相反，我更希望看到另一种新的重大理论发展。从消费者的角度来看，基于参数模型的贝叶斯和频率论分析实际上非常相似。然而，贝叶斯非参数理论尚处于起步阶段，它与频率论非参数理论的联系几乎不存在。我希望有一个更加彻底和成功的贝叶斯非参数理论，从而对无限维问题有更深入的理解。也许必须援引全新的原则来补充 Fisher、Jeffreys、Neyman 和 de Finetti-Savage 的原则。如果发生这种情况，那么越来越非参数化的未来原则上不会将我们推向 Efron 三角形的频率论顶点。相反，贝叶斯推理将继续发挥其启发性的基础作用，重要的新方法将得到发展，甚至可能证明频率派方法可以被视为成熟贝叶斯方法的捷径替代品，具有真正的、深刻的和详细的意义。

评论

Ole E. Barndorff-Nielsen[3]

[3]Ole E. Barndorff-Nielsen 是奥胡斯大学数学研究所理论统计学教授，地址：8000 AARHUS C，丹麦（电子邮件：oebn@mi.aau.dk）。

很高兴阅读 Efron 教授的这篇内容广泛、深思熟虑且勇敢的论文。我同意论文中提出的大部分观点，这篇论文的讨论由一些分散的评论组成，大部分是对论文中提到的内容的补充。

(1) 本文对费舍尔理论在下个世纪的影响的展望中，一个可能重大的遗漏是，没有讨论统计推断的思想和方法在量子力学中可能发挥的作用。这些思想和方法可能会变得越来越重要，特别是随着实验技术的发展，可以研究非常小的量子系统。

此外，现在物理文献中已经出现了大量关于 Fisher 信息的量子类似物以及与 Amari 类似的统计微分几何相关结果的成果。

(2) 强调 “伪似然” 的重要性似乎也很重要，即部分或全部数据和部分或全部参数的函数在很大程度上可以被视为真正的似然。20 世纪下半叶许多最卓有成效的进展都围绕着这种函数。

(3) 我对最优性的看法可能与布拉德利 • 埃夫隆的看法有些不同，我发现对最优性的关注在很大程度上损害了统计发展。问题和危险源于对最优性的定义过于狭隘。一个突出的例子是，过分强调有限的最优性概念可能会偏离科学探究的一般原则，那就是奈曼-佩森检验理论大厦的上层。

一般而言，如果 “最优” 程序不能自然地与全面综合的科学研究方法相契合，那么，即使它 “最优” 又有什么意义呢？

然而，如果我没看错的话，埃夫隆教授和

我对此并不强烈反对。我特别想到的是他提到的 “合理妥协精神”。

(4) 关于在存在许多干扰参数的情况下的近似置信区间和近似基准分布的陈述（第 8.1 节末尾 “置信密度”），我想知道如何将其与 Fisher 关于多参数分布下的基准推断的想法的许多反例相协调。

(5) 关于模型选择，我认为可以发展具有费舍尔精神的技术，尽管这在很大程度上仍有待完成。

(6) 从历史上看，贝叶斯统计学和奈曼-皮尔逊-瓦尔德类型理论都没有在丹麦的统计学中占据一席之地，而类似于费舍尔的思想一直很盛行。

事实上，这一传统可以追溯到 19 世纪后期，尤其是蒂勒 (Thiele)，他早在费舍尔 (Fisher) 之前就开始从事方差分析工作 (参见 Hald，1981)。

评论

D. V. Hinkley [4]

以这篇论文为借口，我在时隔 15 年后重新阅读了《论文集》（Bennett，1972），并再次被其文笔所震撼——通常坚定而无礼，但常常充满幽默感。

费舍尔的工作无疑影响了我对统计学原理和实践的看法，尤其是在 20 世纪 80 年代初期。但我认为，费舍尔大部分工作的生物学背景使他错误地否定了贝叶斯定理作为一种潜在工具的价

[4]D. V. Hinkley 是加利福尼亚大学统计与应用概率系教授，加利福尼亚州圣巴巴拉 93106-3110（电子邮件：hinkley@pstat.ucsb.edu）。

值。阅读费舍尔 (1929) 的两页笔记很有启发性，它回答了学生的建议，即将方差分析等正态理论方法扩展到非正态数据。费舍尔说：“我从未发现生物学工作中会因为变异的不完全正态性而出现困难，尽管我经常检查数据以找出这种特殊的困难原因……这并不是说偏离 [正态理论方法] 在某些技术工作中可能没有实际应用……”（人们不禁要问，费舍尔会如何看待数学金融！）。我认为费舍尔认为抽样分布是具体的，在逻辑上不同于表征先验分布的较弱不确定性，因此与贝叶斯定理的应用不相容。费舍尔关于抽样模型的观点仍然被广泛接受，尤其是在那些使用非贝叶斯方法的人中，因此模型不确定性在几乎所有的统计教育中都不是问题（但在某些科学中并非如此）。但它可能最终成为统计学中最重要的主题之一。请注意，模型不确定性是模型选择的对立面，我们开始使用非费舍尔思想很好地理解它。

其他讨论者可能会对埃夫隆对费舍尔的解释发表评论，因此我只想指出他可能夸大了矛盾。例如，在适当的情况下，随机化推理可以通过适当的设计来调节 - 就像骑士移动和对角正方形一样；参见 Yates (1970, 第 58 页) 和 Savage 等人 (1976, 第 464 页)。事实上，埃夫隆 (1971) 和 Cox (1982) 的结合似乎是典型的费舍尔主义！（类似的想法延伸到重采样-引导-交叉验证分析，但尚未普及。）

但是，Fisher 对统计学的未来可能产生怎样的影响呢？对于 bootstrap，我不确定。当然，参数 bootstrap 涉及使用计算机进行 Fisher 的计算，这种做法是无害的。参数 BC_a bootstrap 方法 (Efron, 1987) 的巧妙理论基础几乎可以由 Fisher 本人编写。超越简单标准误差和标准化估计（用于置信区间或检验）的想法源自 Fisher；参见 Fisher (1928)，在提到样本相关性的传统标准误差时，Fisher 打趣道：“这是 ‘学生必须知道，但只有傻瓜才会使用’ 的事情之一。” 尽管如此，基于可能性的替代发展似乎更接近 Fisher 方法的延续。（所有这些方法都需要进一步发展才能变得容易和广泛使用。）

一个主要问题是非参数引导方法是否具有或将具有 Fisher 的影响。我认为没有。当然，当我们进入模型选择、高度非参数回归（如回归树 (Ripley, 1996)）时，我们更多地处于 Tukey 而非 Fisher 的领域。Fisher 的工作中似乎没有出现纯粹的经验验证和评估，可能是因为无法进行实际计算。

随着元分析思想和方法的传播，经验贝叶斯估计或随机效应建模这一主题变得越来越重要。套用埃夫隆（第 10 节）的话，一个亚群中另一个亚群的信息 “没有明确的费舍尔解释”。但费舍尔肯定会以合理的方式解决这个问题，所以我们需要回过头来阅读费舍尔的作品，弄清楚如何解决这个问题。

Fisher 的主要贡献可能是关于设计的想法，以避免实验结果出现偏差并能够计算出可靠的不确定性度量。与这些相比，关于置信区间是否为贝叶斯的细微之处似乎相对不重要。Fisher 确实为概率在统计学中的作用的讨论做出了宝贵的贡献，其中大部分由 Lane (1980) 进行了有益的调查。就 Fisher 的理论工作而言，尤其是 1925 年和 1934 年的杰作，我一直觉得 Fisher 可能无意中为我们做好了贝叶斯方法论可接受未来的准备。

评论

D. A. S. Fraser[5]

很高兴看到布拉德的深思熟虑和富有洞察力的概述，其中高度赞扬了费舍尔对当前和未来统计学的贡献。布拉德提到了他的评论中未涉及的大量领域，我们的损失是这些领域（例如随机化和实验设计）被忽略了；它们也可能是费舍尔的主要贡献之一，尽管在当前的实践中被忽视了。我强烈支持布拉德对费舍尔贡献的积极认可，并补充一些进一步的认可意见。

布拉德指出，“在布拉德接受教育的时候，费舍尔在美国学术统计学中的地位已经降到了一个相当小的水平。”这似乎是一个相当轻描淡写的说法，尤其是在 1961-1962 年布拉德在斯坦福大学开始他的统计学研究生学习的时候。1961 年秋天，心理学家兼统计学家西德尼 • 西格尔 (Sidney Siegel) 在斯坦福统计学研讨会上发言，描述了他如何绘制应用问题的似然函数图，以深入了解感兴趣的参数，这在现在被视为统计分析中的一个合理步骤。当时，统计学系的教师普遍反对西格尔这样做，认为这不可能也不应该这样做，这是对费舍尔思想的公开拒绝，西格尔感到自己作为统计学家的信誉被剥夺了。布拉德还指出，同年秋天，“费舍尔在斯坦福医学院发表了演讲”。那年秋天，我正在斯坦福统计系参观，确实注意到了费舍尔的演讲。统计系教师缺席费舍尔的演讲是当时费舍尔影响力的一个体现，也许是他影响力的最低点。布拉德做了很多工作来纠正这一问题，并赞扬我们从费舍尔那里继承的财富。

我更喜欢对“频率论者”做出与“竞争哲学……：贝叶斯派；奈曼-瓦尔德频率论者；费舍尔派”略有不同的解释。“频率论者”似乎是指仅由统计模型给出的概率的解释，因此同时涵盖了费舍尔派和决策理论哲学。贝叶斯统计增加了先验概率作为其显著特征。但即使在这里，当我们以“假设”为基础暂时扩大模型来看待这些增加时，这种区别现在可能变得模糊，就像我们经常应用于模型本身的“假设”基础一样。

布拉德称经验贝叶斯“不是费舍尔参与的话题”。然而，它似乎与费舍尔的观点非常一致，事实上，费舍尔应该被视为后来被称为经验贝叶斯的创始人。在 1956 年出版的《统计方法与科学推断》（第 18f 页）一书中，费舍尔考虑了所研究动物的遗传起源，并在初始统计模型中附加了有关该动物起源的理论经验概率。这就是我们现在所说的纯粹的经验贝叶斯，应该归功于费舍尔。从另一个角度来看，这只是扩大了统计模型，以在某种意义上提供适当的建模。

在提到费舍尔解决问题的“手段”时，布拉德有时会使用看似贬义的术语“技巧”。这些手段中的大多数在推出时确实看起来像是技巧，但现在几乎不再是了。也许我们都渴望有更多这些“技巧”作为未来手段的指南。

记录了两个公式 (10) 和 (11)，用于表示给定 [一个] 辅助项“A$ 的 MLE $\hat{\theta}$ 的“条件密度”$f_\theta(\hat{\theta} \mid A)$。这些公式使用了神秘的符号，如果不加以澄清以表明对 $\hat{\theta}$ 的其他依赖关系，则从技术上讲是错误的。上下文假设 $\mathbf{x}$ 等同于 $(\hat{\theta}, A)$，因此密度和可能性的形式为 $f_\theta(\hat{\theta}, A)$ 和 $L(\theta; \hat{\theta}, A)$；然后，放大的公式将显示为

[5]D. A. S. Fraser 是加拿大多伦多大学 Sidney Smith Hall 统计学系教授，邮编：M5S 3G3（电子邮件：dfraser@utstat.toronto.edu）。

$$f_\theta(\hat{\theta} \mid A) = c\frac{L(\theta;\hat{\theta},A)}{L(\hat{\theta};\hat{\theta},A)}$$

$$f_\theta(\hat{\theta} \mid A) = c\frac{L(\theta;\hat{\theta},A)}{L(\hat{\theta};\hat{\theta},A)} \cdot \left\{ -\frac{d^2}{d\theta^2}\log L(\theta;\hat{\theta},A)\Big|_{\theta=\hat{\theta}} \right\}^{1/2}$$

对于位置和一般情况（具有适当的精度）。在位置模型情况下

$$f_\theta(x;\theta) = g(\hat{\theta}-\theta \mid A)h(A)$$

（10）的放大版本则为

$$c\frac{g(\hat{\theta}-\theta \mid A)}{g(0 \mid A)}$$

它重现了条件密度，除了一个常量。相比之下，给定的公式 (10) 从字面上理解，将 $L(\theta)$ 描述为 “作为 θ 的函数，其中 $\mathbf{x}$ 固定”，在观测值 $\left(\hat{\theta}^0, A\right)$ 处给出

$$c\frac{g\left(\hat{\theta}^0-\theta \mid A\right)}{g\left(\hat{\theta}^0-\hat{\theta} \mid A\right)}$$

这是仅针对观测数据点 $(\hat{\theta}, A) = \left(\hat{\theta}^0, A\right)$ 的 MLE 密度近似值。这似乎是一个非常技术性的观点，但这些公式经常被引用为 “计算 $f_\theta(\hat{\theta} \mid A)$ 的方法”。实际上，它们很少用于这样的计算，因为它要求似然函数在 $(\hat{\theta}, A)$ 处可用，其中 A 固定，而 $\hat{\theta}$ 变化，而这需要显式辅助和大量计算。

对于位置模型情况，Fisher 有一个公式，它仅根据观察到的可能性 $L^0(\theta) = L\left(\theta;\hat{\theta}^0, \mathrm{A}\right)$ 给出 $\hat{\theta}$ 的条件密度：

$$f_\theta(\hat{\theta} \mid A;\theta) = cL^0\left(\theta-\hat{\theta}+\hat{\theta}^0\right)$$

其扩展可用于转换模型。

布拉德提到，“这个神奇公式可用来生成近似置信区间，其准确度至少达到二阶。” 事实上，对于连续模型，三阶置信区间具有广泛的普遍性，但需要额外的理论来提供近似辅助。

我特别欢迎 Brad 对基准方法前景的积极看法。作为一名在基准计算具有良好常规属性（变换组环境）或有助于提高重要性值准确性（使用基准消除干扰参数）的环境中广泛工作过的人，我发现在其他地方看到乐观情绪令人欣慰。当然，典型的统计学家认为基准是错误的，但在大多数情况下，他也不熟悉细节或与置信度方法的重叠。

对于变量维度已降低到参数维度的情况，基准和置信度通常都使用枢轴量。置信度程序选择枢轴空间上的 90% 区域，然后反转到参数空间以获取置信区域；相比之下，基准反转到参数空间，然后选择 90% 区域。基准问题源于第二个程序中增加的普遍性。当然，任何观察到的 90% 基准区域都有相应的 90% 枢轴区域，因此当然是该枢轴区域的 90% 置信区域。有争议的问题出现在模拟和重复中。这些重复是否应该始终具有相同的参数值？这不是应用的现实！还有其他选择。我确实赞赏布拉德大胆反转枢轴分布并获得置信密度；基准耳语运动对这个职业的限制太多了。

评论

A. P. Dempster [6]

布拉德利 • 埃夫隆的演讲内容丰富多彩，读起来很有趣。布拉德慷慨而准确地将费舍尔的重要数学基础归功于 “频率学派” 的兴起，该学派的框架显然深深植根于布拉德的心灵和工作中。与此同时，我怀疑他是否充分了解费舍尔的思想，以公正地评价费舍尔对推理逻辑的杰出贡献。对费舍尔 • 奈曼争议的平衡解读表明，20 世纪统计学的历史并不是一条从费舍尔到奈曼，最终到现代 “频率学派” 统计学的线性路径，其主要挑战者是 “主观” 或 “客观” 类型的 “贝叶斯主义”。引号中的陈词滥调需要从根本上澄清和定义，阐明其在实际实践中的科学作用。

Fisher 的推理概念建立在抽样分布与样本数据的解释。样本数据是频率数据，抽样分布具有自然频率解释。但频率的这些作用对于统计学的任何学科观点都是基础，远不能使费舍尔成为奈曼意义上的 “频率主义者”。费舍尔旨在描述数据中的信息，而奈曼则选择了一种理论，该理论指导统计学家在长期运营特征的基础上选择被视为竞争的程序。“奈曼-瓦尔德” 理论提供了有用的见解，但却创造了一种枯燥的实践观点。在选择和应用了程序之后，人们如何思考分析后的不确定性？费舍尔对显着性检验、可能性和基准区间进行了解释，无论它们的优点和缺点是什么，都可以正面回答这个问题。

要理解 Fisher，就需要理解，在实践中，概率决定了特定情况的 “明确规定的不确定性状态”（Fisher，1958 年），这种观点可以定义为描述一种形式上的主观性，它与传统的科学客观性观点相辅相成，而不是相矛盾。在报告结果后，按照 Fisher 的方式解释显著性检验的 p 值（Fisher，1956 年，第 39 页），需要对这种形式上的主观概率进行 “事后” 评估（Dempster，1971 年）。早在 1935 年，Fisher 就认识到 “置信区间” 只是 “另一种说法，即通过某种显著性检验，某些类型的假设可能性将被拒绝，而其他类型的假设可能性则不会被拒绝”（Bennett，1990 年，第 187 页）。正如布拉德所说，人们可能 “普遍认为” 基准概率是费舍尔 “最大的错误”，但不应基于奈曼激烈辩论中提出的简单论点，例如 “相互矛盾的断言的集合，而不是数学理论”（奈曼，1977 年，第 100 页）。解释基准概率与解释贝叶斯后验概率一样，都具有对形式主观概率的预期后分析预测解释，这种解释取决于接受详细的先验假设，特别是在基准论证的情况下，已确定的关键量的分布与观察结果无关，因此不受观察结果的影响。基准推断和贝叶斯推断的成立与否取决于对假设的模型和独立性的逐案判断。

在 20 世纪 50 年代末的吸烟与肺癌争论中，费舍尔是否是 “失势的费舍尔主义者”？他提醒人们注意误导性选择偏差可能会影响观察数据的因果推断，这当然是正确的。费舍尔的创新思想 “如随机化推断和条件性……相互矛盾” 吗？如果人们接受事后推理和预测推理是互补活动，那么就不会。条件性是是互补的活动。条件性对于估计很重要，但对于解释随机化测试的结果无关紧要。此外，在 Fisher 看来，条件性并不是贝叶斯统计的总体原则。事实上，选择性条件对 Fisher 来说是避免普遍服从贝叶斯的关键。Fisher 显

[6]A. P. Dempster 是哈佛大学统计学系教授，马萨诸塞州剑桥 02138（电子邮件：dempster@stat.harvard.edu）。

然没有成功，但他解决了比 Neymanian 理论试图解决的更深、更难的问题。我认为我们应该淡化来自自我限制的理论观点（无论是频率论者还是贝叶斯论者）的简短批评。然而，我不想抗议太多。Brad 和我都同意努力让 Fisher 从美国学术统计学的相对默默无闻中复活。

总结

Bradley Efron

几乎不可能要求有更清晰的讨论，或更合格的讨论者，这让我几乎没有什么可以重新加入的。演讲本身写得很快，只用了几周时间，几乎没有修改就出版了。我担心，仔细的修订会变得太过谨慎，并会因为试图涵盖所有费舍尔的依据而失去其力量。评论中每一条都添加了我省略的重要思想，但彼此之间几乎没有重叠。费舍尔的世界一定是一个非常高维的世界！让我在这里结束，只对评论做出一些简短的反应。

(1) 考克斯教授评论的第 (3) 段涉及费舍尔对概率的定义，该定义取决于可识别子集的缺失。这是统计论据与实际决策相关性的关键点，但这并不是一个容易应用的标准。作为类比，我喜欢想象一片混合着橙色和白色罂粟花的田地，橙色罂粟花的比例代表感兴趣事件的概率。也许橙色比例从东到西、从北到南或从一个角落到另一个角落呈对角线增加。基于逻辑回归或 CART 的模型构建程序相当于系统地搜索该领域的不均匀性。

我们都习惯于根据这些程序的输出，可能表示橙色罂粟花出现在田野中心点的估计概率为 0.20，但这个“概率”与 Fisher 的定义有什么关系？不同的模型，相当于对田野进行不同的划分，可能会给出不同的概率。我的问题不是可识别性，这显然是一个重要的想法，而是它应用于统计人员必须选择哪些子集可能被识别的上下文中。

(2) 卡斯教授讨论了统计哲学中最深奥的问题之一：为什么频率论计算通常有用且引人注目，即使对于那些喜欢贝叶斯范式的人来说也是如此？费舍尔的作品为这个问题提供了迄今为止最好的答案，但我们距离找到令人满意的解决方案还有很长的路要走。卡斯通过贝叶斯非参数方法提出了另一种希望，以弥合这一鸿沟。走这条路需要我们解决另一个深奥的问题，即如何将无信息的先验分布放在高维（或无限维）参数空间中。

(3) 在同一点上，Barndorff-Nielsen 教授询问我们如何将基准推断与 James-Stein 定理等高维估计结果相协调。对于公式 (1) 的多变量版本 $x \sim N_K(\theta, \sigma^2 I)$，原始基准论证似乎表明 $\theta \mid x \sim N_K(x, \sigma^2 I)$，如果我们有兴趣估计 $\|\theta\|^2$，这是一个糟糕的答案。在这种情况下，置信密度方法可以给出合理的结果。我 1993 年的 Biometrika 论文与 Kass 一样认为，这是一种使用频率方法辅助贝叶斯计算的方法。

(4) Hinkley 教授指出，Tukey 式数据分析仍具有持续的生命力。这种工作方式最纯粹的形式是统计学，没有概率论（例如，参见 Mosteller 和 Tukey 1977 年出版的《数据分析与回归》一书），因此我无法将其放在第 11 节的统计三角中的任何位置。这当然是

我的图片的错误，而不是 Tukey 的错误。像神经网络这样的问题驱动领域通常从大量纯数据分析开始，然后逐渐适应统计理论

(5) 我很感谢弗雷泽教授提出了一个更易懂的魔术公式版本。这是演讲中 “避免技术细节” 几乎无法连贯的地方。“技巧” 在我的词汇表中是一个积极的词，反映了加州理工学院的教育，我只希望我能想出更多费舍尔级别的技巧。费舍尔统计是更多 Fisher 级别的技巧。Fisher 统计是统计哲学问题：为什么频数在 20 世纪 60 年代早期的斯坦福大学，生物统计学并非完全缺失：它在医学院蓬勃发展，当时林肯 • 摩西 (Lincoln Moses) 和鲁伯特 • 米勒 (Rupert Miller) 负责生物统计学培训项目。

(6) 我抵制了将讨论者置于统计三角中的冲动，也许是因为邓普斯特教授将我置于比我感觉更不舒服的频率论者角落。邓普斯特强调了一个重要观点：与奈曼的理论相比，费舍尔的理论更侧重于不确定性的后数据解释。这几乎以纯贝叶斯形式出现，带有基准推断，但通常表达得不那么正式，就像他对显着性检验 p 值的解释一样。（参见 1954 年版《研究人员统计方法》第 20 节。）

如果不讨论奈曼，就不能考虑费舍尔，而且他很可能在这样的讨论中被认为是坏人。这是最不公平的。用卡斯的比喻来颠倒一下，奈曼扮演的是尼尔斯 • 玻尔，而费舍尔扮演的是爱因斯坦。没有人能比得上费舍尔的统计推断直觉，但即使是最强大的直觉有时也会误入歧途。奈曼-瓦尔德决策理论是一项英勇的、在很大程度上成功的尝试，它将统计推断建立在可靠的数学基础上。正如巴恩多夫-尼尔森指出的那样，过多的数学可靠性可能会让人感到乏味，但过多的直觉可能会变成神秘主义，人们可以理解奈曼对费舍尔有时德尔菲式的基准声明的失望。

(7) Hinkley 和 Dempster（以及讲座上的其他人）质疑随机化推理是否真的与条件性原则相矛盾。我不得不承认，我自己也曾使用过矛盾的两端，但并没有感到太大的内疚，但逻辑上的不一致似乎仍然存在。随机选择的实验设计难道不是辅助的吗？我们不应该以此为条件，而不是根据其随机性进行推理吗？Dempster 的说法 “条件性对于估计很重要，但对于解释随机化测试的结果无关紧要” 似乎只是重述了这个问题。我们以随机选择的样本大小 n 为条件（借用 Cox 教授的著名例子），无论数据是用于估计还是测试。

(8) 将主观和客观贝叶斯主义混为一谈简化了我的陈述，但可能过于危险。这引起了第 11 节中引用的 Lehman 教授的反对意见。也许最好遵循 Cox 教授对正方形而不是三角形的偏好。

统计哲学最好是少量吸收。讨论者遵循了这条规则（即使我没有），写了六篇精彩的短文。我感谢他们，感谢 COPSS 邀请我做 Fisher 讲座，感谢《统计科学》的编辑们安排这次讨论。

补充参考资料

Bennett, J. H. (1972). Collected Papers of R. A. Fisher. Univ. Adelaide Press

BENnETT, J. H., ed. (1990). Statistical Inference and Analysis. Selected Correspondence of R. A. Fisher. Oxford Univ. Press

Cox, D. R. (1982). A remark on randomization in clinical trials. Utilitas Math. 21A 245-252. (Birthday volume for F. Yates.)

DEMPSTER, A. P. (1971). Model searching and estimation in the logic of inference. In Foundations of Statistical Inference (V. P. Godambe and D. A. Sprott, eds.) 56-78. Holt, Rinehart and Winston, Toronto.

Efron, B. (1971). Forcing a sequential experiment to be balanced. Biometrika 58 403-417.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion). J. Amer. Statist. Assoc. 82 171-200.

Fisher, R. A. (1928). Correlation coefficients in meteorology. Nature 121712.

FISHER, R. A. (1929). Statistics and biological research. Nature 124 266-267.

FishER, R. A. (1935). Design of Experiments. Oliver and Boyd, Edinburgh.

Fisher, R. A. (1956). Statistical Methods and Scientific Inference. Oliver and Boyd, Edinburgh. (Slightly revised versions appeared in 1958 and 1960.)

FishER, R. A. (1958). The nature of probability. Centennial Review 2 261-274.

Gelman, A., Meng, X.-L. and Stern, H. (1996). Posterior predictive assessments of model fitness (with discussion). Statist Sinica 6 773-807.

HALD, A. (1981). T. N. Thiele's contributions to statistics. Internat. Statist. Rev. 49 1-20.

LANE, D. A. (1980). Fisher, Jeffreys and the nature of probability. R.A. Fisher: An Appreciation Lecture Notes in Statist. 1. Springer, New York.

Mahalanobis, P. C. (1938). Professor Ronald Aylmer Fisher. Sankhyā 4 265-272.

Neyman, J. (1977). Frequentist probability and frequentist statistics. Synthese 36 97-131.

RiPley, B. D. (1996). Pattern Recognition and Neural Networks. Cambridge Univ. Press.

Savage, L. J. (1976). On rereading R. A. Fisher. Ann. Statist. 4 441 − 483

YATES, F. (1970). Experimental Design. Selected Papers of Frank Yates, C.B.E., F.R.S. Griffin, London.