

费希尔《实验设计》统计学思想探析

胡 勇^{1,2}

¹西北大学科学史高等研究院, 陕西 西安

²陕西省文化遗产数字人文重点实验室, 陕西 西安

收稿日期: 2026年7月6日; 录用日期: 2026年7月28日; 发布日期: 2026年8月7日

摘 要

费希尔1935年出版的《实验设计》被公认为现代实验统计学的奠基之作。本文以“女士品茶”这一经典案例为切入点, 从“实验作为认识工具”的视角重新审视费希尔的统计学思想。研究表明, “女士品茶”不仅是随机化检验与显著性检验的直观示范, 更是费希尔对其“归纳推断逻辑”的理论演绎。在这一案例中, 随机化程序、零假设检验与确切的概率计算三者构成一个自洽的认识论闭环, 揭示了费希尔实验设计思想的深层结构: 以随机化消除系统性偏倚, 以重复估计误差, 以显著性检验实现“从数据到推断”的归纳逻辑。这一思想途径奠定了当代科学实验的方法论基础, 并深刻影响了波普尔以来的科学哲学发展。

关键词

费希尔, 实验设计, 女士品茶, 统计学

An Exploration of the Statistical Thought in R. A. Fisher's *The Design of Experiments*

Yong Hu^{1,2}

¹Institute for Advanced Studies in the History of Science, Northwest University, Xi'an Shaanxi

²Shaanxi Provincial Key Laboratory of Digital Humanities for Cultural Heritage, Xi'an Shaanxi

Received: July 6, 2026; accepted: July 28, 2026; published: August 7, 2026

Abstract

R. A. Fisher's *The Design of Experiments*, published in 1935, is universally acknowledged as a foundational work of modern experimental statistics. Taking the classic case of “the lady tasting tea” as the entry point, this paper re-examines Fisher's statistical thought from the perspective of “experiment as an epistemic tool”. The study demonstrates that “the lady tasting tea” serves not only as an intuitive

文章引用: 胡勇. 费希尔《实验设计》统计学思想探析[J]. 统计学与应用, 2026, 15(8): 67-75.

DOI: 10.12677/sa.2026.158180

demonstration of randomization tests and significance tests, but also as a theoretical deduction of Fisher's "logic of inductive inference". In this case, the randomization procedure, null hypothesis testing and exact probability calculation form a self-consistent epistemological closed loop, revealing the deep structure of Fisher's thought on experimental design: eliminating systematic bias through randomization, estimating error through replication, and realizing the inductive logic of "from data to inference" through significance testing. This intellectual approach laid the methodological foundation for contemporary scientific experiments and exerted a profound influence on the development of the philosophy of science since Karl Popper.

Keywords

R. A. Fisher, Experimental Design, The Lady Tasting Tea, Statistics

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

在统计学史上, 罗纳德·艾尔默·费希尔(Ronald Aylmer Fisher, 1890~1962)占据着独一无二的地位。1935年出版的《实验设计》(*The Design of Experiments*)不仅将分散于农业试验中的经验方法系统化为一套完整的理论体系, 更从根本上重塑了人们对“科学实验”的理解[1]。正如 Robert Lew 所指出的, 费希尔在这部著作中必须定义并捍卫他关于科学实验、统计推断和概率的全新观念, 然后为初学者提供一套实用的操作程序[2]。在这部著作中, 费希尔明确提出了实验设计的三个核心原则——随机化、重复与局部控制。正如 Oscar Kempthorne 等人所阐释的, 方差分析程序的正当性正是通过费希尔引入的随机化原则来确立的, 而他进一步通过区组设计(blocking)和重复(replication)将这一原则加以充分运用[3]。费希尔由此将统计思想从“数据收集后的分析工具”提升为“实验设计的前提框架”, 这一转变深刻地改变了整个科学研究的认识论基础[4]。

围绕费希尔《实验设计》的统计学思想, 国内外学界已积累了相当丰富的研究成果。从整体上看, 相关研究大致可分为三个方向。

第一是数理统计的技术史路径, 聚焦于费希尔在方差分析、最大似然估计、随机化设计等方面的具体贡献。陈希孺在《数理统计学简史》中对费希尔的工作有专门的梳理与评价[5]。国外方面, Rosenberger 指出, 费希尔“与达尔文一样, 是从无到有地创造了科学”, 其最为持久的贡献包括基于似然的估计、方差分析、实验设计原则以及随机化方法[6]。

第二是科学哲学路径, 侧重于分析费希尔与波普尔、内曼-皮尔逊等人在检验理论上的异同。Ian Hacking (1988)曾对随机化在实验设计中起源的哲学背景进行了深入的史学考察[7]。Lehmann 所著的《Fisher, Neyman, and the Creation of Classical Statistics》(2011)以直接通信和原始文献重构了 Neyman-Pearson 理论的创建过程以及费希尔对此的异议[8]。此外, 正如 Rosenberger 所评论的那样, Fisher 主义方法论之所以能够战胜贝叶斯主义和 Neyman-Pearson 决策理论等逻辑上更为严谨的替代方案, 很大程度上是因为它恰好与逻辑实证主义和波普尔证伪主义在科学哲学中的兴起同频共振[6]。

第三是有关“女士品茶”案例自身的专题研究, 这一方向在近年来尤为活跃。萨尔斯伯格(David Salsburg)的畅销著作(*The Lady Tasting Tea*)以这一案例为主线讲述了 20 世纪统计学的革命史[4], 尽管该书以通俗易懂而著称, 但学者也指出其中存在若干史实不准确之处。Joan Fisher Box 所著的《罗纳德·A. 费

希尔：一位科学家的一生》(*R. A. Fisher: The Life of a Scientist*)对“女士品茶”的真实背景进行了考证：品茶的女士并非剑桥大学某位教授的夫人或宾客，而是费希尔在 Rothamsted 实验站的同事、藻类生物学家 B. Muriel Bristol 博士[9]。值得特别关注的是，Johnny van Doorn 等人从贝叶斯视角重访了“女士品茶”实验，通过 Lindley 的批评揭示了 P 值依赖于抽样计划的根本问题[10]。Mugnier 更深入揭示了费希尔精确检验中一个关键的内隐假设——品茶者是在给定概率信息下最小化期望误分类——并指出，如果没有类似假设或明确的备择假设，费希尔对拒绝域的设定标准就缺乏清晰的合理性论证[11]。Hemerik 和 Goeman 则从数学结构上区分了排列检验和随机化检验，指出费希尔的“女士品茶”实验并非排列检验，而是一个“随机化检验”——前者要求排列集合具有代数意义上的群结构，后者则不然[12]。

尽管既有研究在多个层面取得了重要进展，但各条路径之间仍然存在一个显著的空白地带：数理统计的技术史研究偏重算法与方法的线性演进，科学哲学路径关注理论间的逻辑比较，而“女士品茶”的具体案例分析则多停留在教学阐释或个案考证的层面。费希尔实验设计思想中的“科学性”究竟从何而来？随机化仅仅是去除偏倚的操作技术，还是蕴含着更深层的认识论承诺？显著性检验只是计算 P 值的数学程序，还是承载着一种独特的归纳推理逻辑？换言之，在技术史、哲学分析与案例研究三者之间，尚缺乏一个能够将费希尔实验设计思想“作为一个整体”进行把握的理论框架。

基于此，本文提出一个新的研究视角：将费希尔的实验设计思想视为一套“科学实验的逻辑模型”。这一视角的核心在于，费希尔对实验设计的思考并非仅仅服务于“提高统计效率”这一技术目标，而是力图在“承认误差不可避免”的前提下，为科学知识的确立寻找一种严格的推理基础。换言之，费希尔要回答的根本问题是：在无法直接观察到因果机制的实验条件下，我们如何从有限的可靠数据中作出可靠的推断？“女士品茶”案例之所以具有典范意义，不仅因为它以通俗方式呈现了假设检验的原理，更因为它完整地展示了费希尔对上述问题的回答——一套以随机化为起点、以零假设为参照、以确切的概率计算为桥梁的“归纳推断逻辑”。

本文的研究思路如下：首先以“女士品茶”为切入点，还原费希尔对这一案例的完整论述，厘清其实验设计的数学逻辑；其次，从中提炼并阐释费希尔统计学思想的三个核心维度——随机化的认识论功能、零假设的证伪逻辑以及显著性检验的确切性原理；最后，将这一思想范式置于 20 世纪科学方法论变革的语境中，评估其理论贡献与内在张力。

2. “女士品茶”实验设计

2.1. 实验方案与数学逻辑

在《实验设计》第二章中，费希尔以一段简洁的文字描述了著名的“女士品茶”实验[1]：

“A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested... Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received.”

即一位女士声称，通过品尝一杯用牛奶冲泡的茶，她能够分辨出牛奶和茶水添加的先后顺序。我们将考虑设计一个实验来检验这一主张。我们的实验方案如下：准备八杯茶，四杯按一种方式冲泡，另外四杯按另一种方式冲泡，并以随机顺序呈现给受试者。受试者的任务是将这八杯茶分成两组(每组四杯)，尽可能与实际的冲泡方式相符。

这一实验方案体现了费希尔实验设计思想的三项基本原则。第一，重复原则：每种冲泡方式均有四个重复样本，以降低随机误差的影响，并为误差估计提供依据[1]。第二，随机化原则：八杯茶的呈现顺序通过“物理随机装置”(如骰子、牌或随机数表)决定，从而排除了实验者在呈现顺序上的主观选择可能

引入的系统性偏倚[1]。第三，局部控制原则：虽然这一案例中未明确提及，但其隐含的约束——受试者事先知晓两种冲泡方式各四杯，且必须在品尝后将八杯茶分成两个四杯的集合——实际上发挥了类似于“区组”的功能，控制了受试者在分类数量上的自由度。

在此基础上，费希尔进一步给出了精确的概率计算。从八杯茶中任意选择四杯作为一类，共有：

$$C_8^4 = \frac{8 \times 7 \times 6 \times 5}{4 \times 3 \times 2 \times 1} = 70$$

种可能的组合方式。如果受试者完全没有分辨能力(即零假设成立)，则她正确选出全部四杯“牛奶先加”茶的概率为：

$$p(\text{全部正确} | H_0) = \frac{1}{70} \approx 0.0143$$

恰好选对三杯(即仅错一杯)的概率为：

$$p(\text{恰对3杯} | H_0) = \frac{C_4^3 \cdot C_4^1}{C_8^4} = \frac{4 \times 4}{70} = \frac{16}{70} \approx 0.2286$$

费希尔据此指出，如果受试者全部答对，这一结果便可被视为“统计显著的”(相当于 $P = 0.0143$)，因为它发生的概率低于 2%。

为了更清晰地呈现这一概率结构，表 1 汇总了在零假设下所有可能结果及其概率：

Table 1. Probability distribution of results under the null hypothesis in the “Women’s Tea Tasting” experiment
表 1. “女士品茶”实验中在零假设下的结果概率分布

正确分类数 k	组合数计算式	组合数	概率
4 (全对)	$C_4^4 \cdot C_4^0$	1	$\frac{1}{70} \approx 0.0143$
3 (错 1 杯)	$C_4^3 \cdot C_4^1$	16	$\frac{16}{70} \approx 0.2286$
2 (错 2 杯)	$C_4^2 \cdot C_4^2$	36	$\frac{36}{70} \approx 0.5143$
1 (错 3 杯)	$C_4^1 \cdot C_4^3$	16	$\frac{16}{70} \approx 0.2286$
0 (全错)	$C_4^0 \cdot C_4^4$	1	$\frac{1}{70} \approx 0.0143$

从表 1 可以清晰看到，零假设下的概率分布在 $K = 2$ 处达到峰值，呈对称形态。这一对称性源于“两类各四杯”的平衡设计，使得正确分类数的分布以 2 为中心对称。

2.2. 示范“实验”的本质

“女士品茶”案例之所以成为统计学史上最著名的教学范例，不仅仅是因为它通俗易懂。更重要的在于，它以一种极其精巧的方式示范了费希尔对“科学实验是什么”的回答[4]。

在费希尔看来，一个合格的科学实验必须具备三个特征：第一，实验设计必须预先考虑所有可能的结果，并为每一种结果设定明确的解释(即拒绝或不拒绝零假设的标准)；第二，实验必须通过随机化程序来确保推断的“合理依据”(reasoned basis)；第三，实验的结论不应当是“证明零假设为真”，而应当是在零假设未被推翻的条件下继续持有“无效应”的暂定立场[1]。“女士品茶”案例完美地演绎了这三个特征。费希尔在开展实验之前就计算好了所有 70 种结果的概率分布，并明确了判定标准[1]。随机化程序

确保了受试者的判断结果与冲泡顺序之间不会因实验者的主观偏好而产生虚假关联。而零假设(“受试者无辨别能力”)一旦被拒绝,其结论并非“受试者有辨别能力”的绝对证明,而只是在概率意义上排除了偶然因素的干扰——这正是费希尔所说的“每一个实验都可以说仅仅是为了给事实一个机会来证伪零假设”[1]。

事实上,“女士品茶”案例的深远意义还体现在它对后续统计学发展的持续激发上。研究表明,这一案例虽然常被称为“第一个排列检验”(permutation test),但从严格的数学结构来看,它实际上是一个“随机化检验”(randomisation test)[12]。两者的区别在于:排列检验要求排列集合具有群结构,而随机化检验的基础是实验中的物理随机化,其数学推理不依赖于群结构[12]。这一细微的差异揭示了一个重要的史实:费希尔的实验设计思想并非从既有的数学理论出发去改造实验,而是从实验的物理过程中衍生出概率推断的逻辑。换言之,随机化之所以是统计推断的“合理依据”,不是因为它满足某个先验的数学条件,而是因为它是一种可以实际执行的、能够消除系统性偏倚的物理程序——这一洞见至今仍深刻影响着实验设计的思想基础。

3. 费希尔实验设计思想的三重维度

3.1. 随机化的认识论功能

在费希尔的实验设计框架中,随机化居于核心地位。然而,随机化的功能并非像通常理解的那样仅仅是“使实验组与对照组在已知因素上平衡”。费希尔对随机化的论证远比这一常识更为深刻。

首先,费希尔指出,随机化能够同时处理“已知”和“未知”的混杂因素[1]。这一点在其农业试验背景中尤为明显:两块试验田在土壤肥力、排水条件、光照等众多方面都可能存在差异,其中许多差异甚至是实验者完全不知道的。经典的配对或区组设计只能控制已知的变异来源,而对未知的干扰因素无能为力。但随机化通过“抽签”的方式将处理分配到各个实验单元上,使得任何(无论已知还是未知)的系统性差异被期望为零地分散到各处理组中,从而使得观察到的处理效应与随机误差之间的比较成为可能[1]。

其次,更根本的是,费希尔认为随机化是统计推断的“合理依据”(reasoned basis)。在他之前的统计实践中,研究者通常在数据收集完成之后才请统计学家来分析,统计学家的工作仅限于对已经摆在面前的数据进行计算。费希尔彻底颠覆了这一顺序:他认为统计推断的可信度不是来自事后分析的精密程度,而是来自事前实验设计的严谨性——而随机化正是这种严谨性的核心。只有通过随机化程序,我们才能合法地将概率分布“施加”到实验数据上,从而赋予显著性检验以逻辑基础。

正如费希尔本人所强调的,“归纳推断是我们所知道的唯一能使本质上全新的知识进入世界的途径。”[1]而随机化恰恰是这种归纳推断得以在实验科学中实现的核心机制:它将实验者从必须事后为数据寻找概率模型的困境中解放出来,使得概率模型成为实验设计的“出厂设置”而非事后修补。

3.2. 零假设与证伪逻辑

“女士品茶”案例中所采用的零假设检验法,蕴含着一种独特的科学哲学立场。费希尔在《实验设计》第16页写下一句广为引用的格言:

“Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.”[1]

这句话与卡尔·波普尔在1934年《科学发现的逻辑》中提出的“证伪主义”惊人地相似。正如Rosenberger所评论的那样,Fisher主义方法论恰好是在新实证主义与证伪主义在科学哲学中兴起之时出现的,费希尔建议研究者将注意力集中在证伪零假设之上[6]。

然而，费希尔的“证伪”与波普尔的“证伪”存在着一个根本性的差异。波普尔主张的是逻辑上的严格证伪：只要有一个与理论预测不符的观察事实，该理论即可被判定为伪。但费希尔面对的是带有随机误差的实验数据，不可能存在如此严格的“矛盾”。因此，费希尔以概率论来“软化”证伪的逻辑要求：如果零假设为真，那么观察到当前结果(或更极端结果)的概率(即 P 值)很小，我们就可以说实验“否定”了零假设——但这里所谓的“否定”从来不是绝对的，而是一种基于概率的、可错的判断。

这一“概率化证伪”的逻辑深刻地塑造了 20 世纪的科学方法论。正如费希尔本人在《实验设计》序言中所言，“对简单而标准化的统计程序的清晰把握……将在很大程度上阐明实验的原理；但程序本身只是手段，其角色是满足可靠而可理解的实验设计的要求，并为无歧义的解释提供机制。”^[1]费希尔的意义并非仅仅在于他提供了 P 值这一技术工具，而在于他为科学家如何在不确定条件下进行理性推断提供了一套完整的操作程序：

“设定零假设 H_0 → 通过随机化设计收集数据 → 计算 P 值 → 若 P 值足够小则拒绝零假设 H_0 ”

这套程序既尊重了经验科学对“检验”的追求，又接纳了误差不可避免的认识论前提。

3.3. 显著性检验的确切性

在理解费希尔统计学思想时，一个不可回避的问题是他与耶日·内曼(Jerzy Neyman)和埃贡·皮尔逊(Egon Pearson)之间的深刻分歧^[8]。虽然两者都使用显著性检验，但对检验的性质和目的有着截然不同的理解。

内曼-皮尔逊框架将假设检验理解作为一种决策理论：研究者设定零假设和备择假设，预先确定第一类错误和第二类错误的可容忍水平，然后在给定的显著性水平上做出“接受”或“拒绝”的决定^[8]。这一框架具有很强的实用性，尤其适用于质量控制、医学检验等需要明确行动准则的领域。

费希尔则持有截然不同的立场。他坚决拒绝将显著性检验简化为一种“接受或拒绝”的二元决策规则。对他而言， P 值是一种“归纳推断的测度”——它衡量的是数据与零假设之间的不一致程度，而不是一个预先设定的决策边界^[1]。在“女士品茶”案例中， $P=0.0143$ 是一个具体的、由实验设计本身的组合结构所决定的数字，它仅仅告诉研究者：如果零假设为真，这样的结果有多么罕见。至于这一罕见程度是否足以构成“证据”，费希尔认为这取决于研究者的判断和具体的科学语境，而非机械地套用某个阈值。

正如 Mugnier 所准确指出的，费希尔精确检验的一个关键特征是它建立在品茶者“给定固定概率信息下最小化期望误分类”这一内隐假设之上——这一发现揭示了 Fisher 检验的深层结构：“没有类似假设或明确的备择假设，费希尔对拒绝域的设定标准就缺乏清晰的合理性论证。”^[11]这一区分在根本上涉及对“统计推断”本性的理解。内曼-皮尔逊试图将统计推断数学化为一种行为理论，而费希尔始终坚守将其理解作为一种科学研究方法论。前者追求的是决策规则的最优性，后者追求的是从实验设计中自然生成的、未经人为调参的确切概率。“女士品茶”之所以是一个完美的范例，正在于它展示了一种无需备择假设、无需预先设定显著性水平就能得出确切概率的检验逻辑——这正是费希尔范式的独特魅力所在。

4. 哲学预设与贝叶斯批判

费希尔从随机化实验到科学结论的推断链条，其表面上的自治性掩盖了若干深层哲学假设。揭示这些预设，并考察其他学派(尤其是贝叶斯学派)的批判，不仅不会削弱费希尔思想的地位，反而能通过承认其复杂性和争议性，更客观地界定其适用范围与理论边界。

4.1. 三个关键预设

费希尔的“设计-执行-推断”逻辑至少依赖于以下三个不可回避的哲学假设：

第一，随机化作为“概率基础”的合法性问题。费希尔将物理随机化视为赋予数据概率解释的合法依据。然而，这一论点预设了一个认识论立场：实验者主动引入的随机性可以替代对误差结构的实质性知识。一个可操作的物理程序何以能够产生认识论上的规范性力量？费希尔对此并未给出系统的哲学辩护，而是将其作为一个不言自明的实验实践基础。

第二，“无效应”零假设的认识论优先性。费希尔选择“受试者无辨别能力”作为零假设，并赋予其被“证伪”的目标地位。这一选择并非逻辑上的必然，而是一个策略性的认识论决定——它使得研究者始终处于“试图推翻自己假设”的严苛立场。科学进步来源于大胆猜测和严厉检验。费希尔的零假设检验同样是建立在这一科学进步观之上的规范性选择。

第三，从P值到科学结论的“归纳跳跃”。即使P值被正确计算，从“数据与零假设不一致”到“实验效应存在”这一科学结论之间，仍然存在一个无法被纯粹概率覆盖的推理空间。费希尔本人承认这一判断依赖于“研究者的判断和具体的科学语境”，但并未为这一判断提供明确的逻辑规则。

4.2. 贝叶斯学派的系统批判

贝叶斯学派从逻辑基础出发，对费希尔的推断框架提出了更为系统的批评。其核心论点为：费希尔的P值并非研究者真正关心的量——研究者真正想知道的是“在观察到当前数据后，零假设为真的概率”，即 $P(H_0|data)$ ，而P值给出的却是 $P(data|H_0)$ 。这两个量在逻辑上不等价，且二者之间的关系依赖于先验概率 $P(H_0)$ ，而费希尔体系刻意回避了对先验的讨论。

van Doorn 等人曾以“女士品茶”为案例，从贝叶斯视角进行了重访分析[10]。他们指出，Lindley 的经典批评在此案例中有着直观的体现：即使在 $P=0.0143$ 这样看似“显著”的结果下，如果研究者事先认为“受试者有辨别能力”的先验概率很低，贝叶斯后验概率 $P(H_1|data)$ 可能仍然不具压倒性。具体而言，在“女士品茶”案例中，假设先验概率 $P(H_1)=0.5$ （即无信息先验），则在观察到“8杯全对”的结果后，后验胜算为

$$\frac{P(H_1|data)}{P(H_0|data)} = \frac{P(H_1)}{P(H_0)} \times \frac{P(data|H_1)}{P(data|H_0)}$$

其中，若假设 H_1 下品茶者能以概率 $1-\varepsilon$ 正确分类（ ε 为很小的犯错概率），则似然比约为 $\frac{(1-\varepsilon)^8}{\left(\frac{1}{70}\right)}$ 。即便

$\varepsilon=0.05$ ，似然比也仅约为 $0.95^8 \times 70 \approx 46.4$ 。若先验胜算为1:1，后验胜算约为46.4:1，对应的后验概率 $P(H_1|data) \approx 0.979$ ，虽低于0.986的P值换算结果但提供了较强的证据支持。若先验更为保守，如： $P(H_1)=0.1$ ，则后验概率仅约为0.84，远低于 $1-P$ 的直观解读。这一分析揭示了费希尔体系对先验信息刻意回避所带来的认识论代价：当背景知识能够提供先验信息时，单纯依赖P值可能会系统性地夸大或低估证据的强度。

然而，贝叶斯批判本身并非费希尔体系的“致命缺陷”，而更应被理解为对同一科学推断问题的不同回答。费希尔执着于从实验设计内部产生概率解释——即“从数据内部说话”——因而拒绝了外部先验信息的介入；而贝叶斯学派则主张将一切可用信息(包括先验信念)纳入概率演算的统一框架。两者的分歧本质上是关于“科学推断应当如何整合信息”的根本性分歧，而非简单的技术路线差异。

4.3. 对费希尔思想的再审视

上述哲学预设与贝叶斯批判提示我们，费希尔的实验设计思想并非一套逻辑上自足完备的“科学推断算法”，而是一种高度精致的、依赖于特定科学价值观的方法论体系。其核心价值在于提供了从不确定性实验中获取可靠知识的操作程序，但其局限性同样明显：当P值被从其完整的“设计-执行-推断”语境中剥离出来、被简化为一个孤立的数字时，费希尔本人所强调的“判断依赖于语境”这一关键环节便消失了——这正是当代统计实践中诸多误用的根源所在。

理解费希尔思想的边界，不是为了否定其贡献，而是为了更准确地把握其适用条件：费希尔的推断逻辑只有在随机化设计被严格遵守、零假设被正确定义、且P值的计算基于预设的实验方案(而非数据窥探后的回溯)时才能保持其认识论上的严谨性。这一认识将我们直接引向对当代“可重复性危机”的分析。

5. 结语

“女士品茶”案例诞生近九十年后，我们依然可以从这一小小的思想实验中汲取丰富的灵感。费希尔的实验设计思想在当代科学实践中的持续生命力，不仅体现在农业试验、医药研发、社会科学实验等广泛领域中随机化对照试验(RCT)的普遍采用，更体现在它所蕴含的认识论价值至今仍在激发学界的讨论。

从技术层面看，费希尔提出的随机化、重复与局部控制三项原则已成为任何严谨实验设计的基本规范，贯穿于从基础科学研究到大规模临床试验的各个领域。从思想史层面看，费希尔的显著性检验——尤其是其零假设证伪逻辑——深刻地影响了波普尔以后的科学哲学发展。

然而，近年来心理学、医学等领域爆发的“可重复性危机”(replication crisis)却迫使学界重新审视：如果费希尔的范式如此严谨，为什么大量已发表的研究结果无法被重现？对这一问题的深入反思揭示了一个令人警醒的结论：当代“可重复性危机”的根源，恰恰在于研究者系统性地破坏了费希尔所构建的“设计-执行-推断”这一完整逻辑链条。具体而言，这种破坏表现为以下形式：

P-hacking (P值操纵)。研究者在数据收集完成后，通过选择性报告、添加或删除协变量、变换变量定义等方式，反复尝试直至得到“显著”的P值。这种做法完全颠覆了费希尔对P值的理解——在费希尔体系中，P值的概率解释依赖于实验方案的预先设定和数据的完整收集；一旦实验方案被事后调整，P值的数学含义便不再成立。P-hacking使得本应由物理随机化保证的概率分布被人为的数据探查所取代，P值不再是“如果零假设为真，观察到当前结果的概率”，而是变成了“如果研究者不断尝试，最终找到显著结果的概率”——这一概率可能远高于名义上的0.05。

HARKing (Hypothesizing After the Results are Known, 事后假设)。研究者在看到数据后才提出研究假设，并将其包装为事先的预测。费希尔的零假设检验逻辑要求假设必须在实验之前明确设定；事后假设破坏了假设检验的时间顺序，使得本应作为“证伪对象”的零假设变成了研究者可以事后裁剪的“靶子”。这一做法使得P值失去了其“证伪”的认识论意义。

上述实践的共同特征在于：它们保留了费希尔显著性检验的“语言”：P值、零假设、显著性水平，却抽空了这一语言赖以成立的认识论基础。从这个意义上说，“可重复性危机”并非费希尔思想的失败，而恰恰是费希尔思想被肢解、误用的后果。

费希尔在《实验设计》序言中表示，对简单而标准化的统计程序的清晰把握……将在很大程度上阐明实验的原理；但程序本身只是手段，其角色是满足可靠而可理解的实验设计的要求，并为无歧义的解释提供机制。这一论述提示我们，费希尔的遗产远不止于方差分析公式或P值的计算方法。它更是一套关于“如何在不确定性的世界中获取可靠知识”的方法论哲学。从这个意义而言，“女士品茶”不仅是一个统计学案例，更是一部关于科学探究之本质的思想小品。当那位女士需要分辨的是茶与牛奶的添加顺序，

当费希尔试图解决的问题是随机误差与真实效应的边界，他们所触及的，实质上是人类知识何以可能的元问题——而这，或许正是数学史研究者从“女士品茶”中能够领悟的最深刻的统计学思想。

致 谢

感谢西北大学科学史高等研究院陈克胜副教授、赵继伟副教授的宝贵建议。

参考文献

- [1] Fisher, R.A. (1935) *The Design of Experiments*. Oliver & Boyd.
- [2] Lew, R. (2014) *Experimental Design*. In: *Wiley StatsRef: Statistics Reference Online*, Wiley.
<https://doi.org/10.1002/9781118445112.stat05892>
- [3] Kempthorne, O. (1980) The Teaching of Statistics: Content versus Form. *The American Statistician*, **34**, 17-21.
<https://doi.org/10.1080/00031305.1980.10482704>
- [4] Salsburg, D. (2001) *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. W.H. Freeman and Company.
- [5] 陈希孺. 数理统计学简史[M]. 长沙: 湖南教育出版社, 2002.
- [6] Rosenberger, W.F. (2026) The 41st Fisher Memorial Lecture from Fisher to CARA: The Evolution of Randomization and Randomization-Based Inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **189**, 497-511. <https://doi.org/10.1093/jrsssa/qnaf002>
- [7] Hacking, I. (1988) Telepathy: Origins of Randomization in Experimental Design. *Isis*, **79**, 427-451.
<https://doi.org/10.1086/354775>
- [8] Lehmann, E.L. (2011) *Fisher, Neyman, and the Creation of Classical Statistics*. Springer.
- [9] Box, J.F. (1978) *R.A. Fisher: The Life of a Scientist*. Wiley.
- [10] van Doorn, J., Matzke, D. and Wagenmakers, E.J. (2020) An In-Class Demonstration of Bayesian Inference. *Psychology Learning & Teaching*, **19**, 36-45. <https://doi.org/10.1177/1475725719848574>
- [11] Mugnier, M. (2024) R.A. Fisher's Exact Test Revisited. <https://doi.org/10.48550/arXiv.2407.07251>
- [12] Hemerik, J. and Goeman, J.J. (2021) Another Look at the Lady Tasting Tea and Differences between Permutation Tests and Randomisation Tests. *International Statistical Review*, **89**, 367-381. <https://doi.org/10.1111/insr.12431>