AgroDesign：面向农业试验的设计感知统计推断框架（Python）

Aqib Gul IBPR，克什米尔谢赫·卡什米尔农业科学与技术大学
maaqib90@gmail.com

arXiv:2603.09041v1 [stat.ME] 10 Mar 2026

摘要

农业试验的统计分析基于随机区组、因子、裂区设计和多环境试验等结构化实验设计。尽管这些方法的理论基础十分扎实，但在现代编程框架中的实现通常需要手动指定统计模型、选择误差项，并对交互作用效应进行主观解释。实验设计与计算实现之间的这一脱节，容易导致误导性的推断和不一致的报告。我们推出 AgroDesign，这是一个以实验设计为统计分析核心规范的 Python 框架。该框架将指定的实验设计直接转换为有效的线性模型，自动识别误差层，开展假设检验和均值分离，检验线性模型的假设，并提供以决策为重点的解释。该框架将固定效应方差分析、层次设计、线性混合模型以及基因型×环境稳定性分析整合到一个声明式框架中。AgroDesign 在农业统计学中的经典设计上得到验证；结果表明，其与传统统计分析保持一致，同时严格执行正确的解释约束，尤其适用于交互作用占主导地位的设计和多层设计。通过将设计语义融入计算过程，该框架减少了由分析师驱动的建模选择，并提升了可重复性。

1 引言

1.1 农业试验中的统计分析

受控试验是农业与生物科学推断的基础。田间和温室试验通常采用有组织的实验布局开展，例如完全随机设计（CRD）、随机完全区组设计（RCBD）、析因设计、裂区试验和多环境试验。这类设计为区分处理变异与环境变异提供了手段，也是提出统计上有效的农艺建议的基础 [2]。与一般回归分析不同，设计实验中的推断不仅依赖线性模型，也依赖随机化设计及相应的误差层（Morris，2010）。在区组设计和层次设计中，每个因素都需要与适当的实验单位进行比较。错误的均方会导致不恰当的 F检验，进而可能得出错误的科学推断 [6, 7]。因此，正确的分析需要确定实验单位、建立适当的方差分析表、比较处理、验证模型假设，并以农艺学术语解释统计结果 [2]。尽管方差分析的理论基础十分扎实，但要正确应用它，需要具备较高水平的专业知识，涵盖

统计建模和实验设计理论。这往往导致结论的有效性取决于分析师的专业水平，而非有据可查的试验设计。随着计算工作流对科学分析的影响日益加深，缺乏对实验结构的正式表示，使实验设计理论与其在软件中的实现之间出现脱节 [3]。即使实验设计相同，同一数据集也可能因模型构建的结构不同而得出不同的统计推断。这表明，实验设计不仅是用于规划和分析的概念框架，还是一种结构化信息，必须在分析过程中明确编码。将实验设计作为一等计算对象，可用于编码随机化结构、限制模型构建，并确保统计推断依据经典方差分析的原则，而非依赖分析师的解读。

1.2 当前计算工作流的局限

当代科学计算环境支持丰富的统计建模，但大多处于以回归为中心的框架中。例如，Statsmodels等Python软件包通过公式接口支持线性建模和方差分析 [8]。尽管这些软件包在数学上自成体系，但仍要求分析师明确指定交互项及相应的误差结构。在裂区设计或多因素研究等层次实验设计中，这项任务尤其容易出错。其他科学计算软件包提供离散的分析功能。数值计算工具包提供假设检验和分布 [9], ，专用软件包则提供多重比较程序（Terpilowski，2019年）。实验设计矩阵生成器可提供设计，但缺乏统计推断。因此，分析过程需要组合互不关联的工具，使分析更加复杂且容易出错。可重复性研究还发现，分析灵活性以及隐含的研究者选择会影响统计结果 [11, 12]。对于设计实验，手动构建模型和事后检验会引入更多自由度，难以保证不同分析师和不同实验之间解释的一致性。在其他生态系统中，也存在面向设计农业试验的领域专用统计软件包，例如R环境中的专用统计平台和软件包。然而，这些软件包往往与用于预处理、数据可视化和机器学习的当代数据科学工作流相互隔离。这会造成重复劳动，并阻碍其与可重复计算研究工作流的整合。在农业统计学中，一直需要专用软件来处理这一问题，而面向设计的分析环境已经做到了这一点。例如，R中的agricolae等专用软件包具备经典试验设计中处理比较与分析的程序 [1]。然而，这些软件包是作为独立的统计软件运行的，而非通用计算工作流的一部分。现代科学计算正将数据处理、可视化和机器学习整合到统一的编程环境中，但实验设计的分析却在该环境之外进行。这意味着科学家必须在不同环境之间切换：先导出数据进行统计分析，再将结果导入以进行后续处理。

1.3 本文的贡献

本研究提出了一种用于统计推断的设计驱动框架，其中实验设计成为分析的核心规范，而非由用户指定的统计公式。实验设计由分析师指定，框架会自动

推导推断所需的统计程序。系统根据实验设计自动构建相应的方差分析模型，识别正确的误差层，区分处理均值，检验模型假设，并提供具有统计学意义的结果。因此，分析直接遵循编码后的随机化设计，而不是分析师指定的建模选择。该框架以名为 agrodesign 的开源 Python 系统实现，该系统将实验设计视为一等计算对象。该系统并非由一组松散耦合的统计检验构成，而是采用严格的推断工作流，将模型构建、假设检验、事后检验和生物学解释串联起来。本研究的贡献如下：

1. 一种用于实验设计的声明式框架，将实验单位和随机化层级纳入计算流程。2. 根据指定的设计，自动推断线性模型结构及相应的误差项。3. 一种受约束的推断流程，结合方差分析、多重检验和假设检验，以尽量减少分析师指定的参数。4. 一个决策分量，将统计结果转化为处理比较和农学解释。5. 通过与模拟数据和实验数据上的标准统计方法进行比较，开展实证验证。

通过将实验结构形式化为可执行的分析规范，所提出的框架能够对农业试验进行可复现且统计一致的分析，同时兼容现代科学计算环境。

2 统计框架

2.1 作为结构化统计模型的实验设计

在经典方差分析中，统计模型并非独立于实验指定，而是由用于生成数据的随机化方案决定。每种实验设计都隐含地指定了：(i) 可估效应集合；(ii) 每种效应对应的适当实验单位；以及 (iii) 用于有效假设检验的分母均方 [2, 5]。在这种情况下，实验设计可视为一种结构化统计对象，而非描述性元数据。设 Y 为通过随机化程序创建的实验单位上测得的响应变量。设计指定了单位层级结构 $U_1 \subseteq U_2 \subseteq \cdots \subseteq U_k$，处理在该层级结构的特定层级上施加。推断的有效性取决于将每种处理效应与同一随机化层级上的变异进行比较。因此，具有相同数据但分配结构不同的两个实验并不相同，且对应不同的统计模型。例如，在处理因子为 T_i 的完全随机设计中，观测值满足

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

其中所有实验单位共享相同的误差方差。在随机完全区组设计中，额外的区组因子 B_j 表示环境异质性：

$$Y_{ij} = \mu + \tau_i + B_j + \varepsilon_{ij} \tag{2}$$

其中，处理效应是相对于区组内变异而非总变异进行检验的。层次设计引入多个误差层。在主区因素 A 和副区因素 B 的裂区试验中，存在两个独立的误差项：

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + n_{k(i)} + \varepsilon_{ijk} \tag{3}$$

其中，$n_{k(i)}$ 表示主区误差，e_{ijk} 表示副区误差。将某因素与e_{ijk} 而非$n_{k(i)}$ 进行检验，会产生错误的F统计量（Steel等人，1997年）。类似地，多环境试验会引入交叉随机因素：

$$Y_{ijk} = \mu + G_i + E_j + (GE)_{ij} + \varepsilon_{ijk} \tag{4}$$

其中，解释取决于交互项 [7]的大小及其显著性。这些例子表明，试验设计决定了统计模型、假设检验以及推断范围。将设计信息用作外部注释，意味着分析师必须手动推断这些约束，而这一过程可能含糊不清且容易出错。本文提出的方法将实验设计形式化为对这一结构的计算表示。设计规范不再要求指定统计模型，而是描述实验单位的嵌套关系以及处理的分配方式。随后，根据这一表示推断统计模型和有效的推断程序。

2.2 基于实验结构的模型构建

根据给定的数据集和明确说明的实验设计，该框架会确定一种可接受的统计分析方法，与第3.1节所述的模型相对应。实验设计不仅规定要拟合哪种线性模型，也规定适用的估计量和假设检验。这样，估计与推断便相互分离：参数按照标准统计方法进行估计，而设计则规定允许进行的推断操作。对于单层实验单位，所述设计规定采用完全随机模型（1）。处理间比较基于残差变异。在设计中指定区组后，则采用随机完全区组模型（2），处理间比较在区组内进行，而非在整个数据集范围内进行。对于析因设计，则自动采用交叉模型（3），并纳入设计所规定的所有交互项，无需分析师另行指定。层级布局会引入多个实验单位。采用式（3）类型模型的裂区设计会估计完整的线性模型，但会依据为各因素定义的误差层进行假设检验。主区效应以主区误差为依据进行检验，而副区效应则以残差误差为依据进行检验。估计方法保持不变，但推断的有效性由随机化方案而非线性模型决定。有些试验同时包含处理的固定效应和环境因素的随机效应。指定随机因素后，分析会采用如下形式的线性混合模型

$$Y = X\beta + Zu + \varepsilon \tag{5}$$

其中，固定效应用于衡量处理表现，随机效应用于衡量环境变异。与比较简单均值不同，处理表现通过最佳线性无偏预测值进行评估，从而得出考虑不同环境和缺乏重复情况的调整估计值。推断的目标也从比较均值转变为预测随机环境中的处理表现。对于由方程（4）建模的多环境试验，基因型与环境的交互作用界定了有效推断的范围。指定的交互结构

随后进一步分解，以描述稳定性和适应性。加性主效应与乘性交互作用分解、Finlay–Wilkinson 回归以及 Eberhart–Russell 稳定性参数，既描述平均表现，也描述对环境的响应一致性。这些推断并非被视为彼此独立的分析，而是同一底层模型的推论，从而确保稳定性分析与表现预测之间的统计一致性。同一结构模型可推广至多个响应变量和分组实验。对于多个性状，针对每个响应变量分别指定相同的基于设计的模型，同时保持相同的解释结构。对于重复年份或地点等分组数据集，数据集被划分为若干子集

$$D = \bigcup_{g \in G} D_g \tag{6}$$

并且，在每个子集中应用实验设计所隐含的相同模型。由此得到的分析构成了一组结构化且可比较的推断结果，因为各组采用的统计程序保持不变。

2.3 层级推断方案

在设计实验中，统计显著性不足以得出有效结论。如何基于处理效应得出结论，取决于拟合数据的模型所具有的层级结构。在析因设计和层次设计中，若存在高阶交互作用，低阶效应可能就没有意义。然而，标准统计软件包会一次性给出所有检验结果，而分析师只能自行得出结论，这可能导致结论前后不一致。本文提出了一种基于拟合模型结构的受约束的推断问题。设拟合模型包含一组效应

$$\mathcal{F} = \{E_1, E_2, \ldots, E_n\} \tag{7}$$

每个效应都对应一个假设检验。将效应的阶数定义为该项所涉及的因子数量。可接受的解释域由统计显著的最高阶效应决定

$$E^* = \arg\max_{E_i \in \mathcal{F}} \{\text{order}(E_i) \mid p(E_i) \leq \alpha\} \tag{8}$$

可接受的解释域由公式 8 中定义的主导效应决定。所有不符合 E^* 的低阶效应均不纳入解释。因此，若交互作用显著，则不应解释边际均值，因为它们是对处理反应的平均值，而这些处理并非同质。相反，推断仅限于联合因素中的简单效应。事后分析仅在可接受区域内进行。令 M(E) 表示效应 E 的估计边际均值。多重比较程序仅限于

$$\mathcal{M} = \{M(E) \mid E \in E^*\} \tag{9}$$

因此，当交互作用占主导时，应在处理组合之间进行比较；只有在不存在更高阶交互作用时，才应比较主效应。这样可以避免基于对不均一处理响应取平均而作出可能无效的比较。上述层级限制可以推广到其他析因模型。在裂区设计中，推断仅限于定义在不同实验单位上的因子。在混合模型中，处理

比较依据预测表现而非原始值进行，因为存在随机环境效应。在多环境试验中，根据基因型与环境交互作用的大小，可以作出总体推荐，也可以作出针对特定环境的推荐。通过将推断规则直接纳入推断过程，分析结果便可根据实验设计和显著性水平确定。分析师定义实验，并根据模型结构得出有效结论。

2.4 假设有效性与诊断约束

上述推断步骤建立在线性模型相关的分布假设之上。要使假设检验和处理比较有效，模型的残差必须满足推导 F统计量和多重比较程序时所依据的假设。因此，需要注意的是，假设检验核查被视为推断过程的一部分，而不是分析完成后的诊断步骤。令 $\hat{}\varepsilon_{ijk}$ 表示拟合模型的残差。对于固定效应分析，允许的推断要求

$$\hat{\varepsilon}_{ijk} \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2) \tag{10}$$

这意味着需要满足两个实际条件：残差正态性，以及各处理组之间的方差齐性。在假设检验中，正态性是 F分布近似有效的必要条件；而同方差性则是分母中的均方成为误差方差无偏估计量的必要条件。违反这些条件并不会阻止估计，但会影响推断的有效性。因此，该框架引入了一个有效性谓词

$$\mathcal{V}(D, \mathcal{E}) = \begin{cases} \text{true} & \text{if model assumptions are not rejected at level } \alpha_v, \\ \text{false} & \text{otherwise.} \end{cases} \tag{11}$$

假设检验和处理比较均以 $\mathcal{V}$ 为条件进行解释。当满足这些假设时，结论被认为在统计上有效。假设检验在与每个被检验效应相对应的误差层上进行评估。在层级模型和混合模型中，对于与每个实验单位相关的残差部分，同样适用。对于裂区设计，假设检验核查与假设检验所采用误差项相关的副区残差有关。对于混合模型，假设检验核查适用于调整随机效应后的条件残差。这样一来，有效性标准取决于决定推断的误差层，而不是整个数据集。通过将假设检验核查纳入推断过程，统计结论取决于模型有效性，而不是被默认接受。因此，分析过程区分了可计算的结果与有效的结论，从而确保基于实验得出的结论与统计推断所依据的概率假设保持一致。

2.5 面向决策的解释

该框架的最后一步是将统计上可接受的结论转化为正式的决策规则。在设计实验中，统计显著性并非得出实际结论的标准；实际结论取决于实验目标以及所推断效应的正式结构。该框架将统计输出视为分析过程中的中间形式，而非分析结果。令 $\mathcal{M}$ 表示第 2.3 节确定的估计均值可接受集合。决策函数δ：$\mathcal{M} \to \mathbb{R}$ 将统计上有效的比较映射到建议集 $\mathbb{R}$。该函数根据实验设计所隐含的模型类型来定义。在固定效应实验中，

决策规则依据估计边际均值对处理进行排序，同时考虑多重比较检验施加的分组约束。属于同一统计组的处理被视为等效，建议则以统计等效最优处理构成的集合来表示，而非单个数值最大值。当交互作用效应占主导时，排序是在各因子组合内进行，而不是依据边际平均值进行。在混合模型分析中，决策依据的是预测表现，而非观测均值。令 $\hat{u}$ 表示由混合模型得出的预测处理效应。建议规则选择使预测响应最大化的处理，并考虑随机环境变异，因此决策依据的是预期表现，而非观测结果。

3 实现

该框架被实现为一个计算系统，用于执行第2节中定义的统计程序。实现并不引入新的估计量，而是确保模型构建的正确性和允许的推断，同时将数值计算交由成熟的统计库完成。因此，该软件充当统计框架的执行层。框架的总体架构如图1所示。该流程分为两个阶段：设计规范和允许的推断。

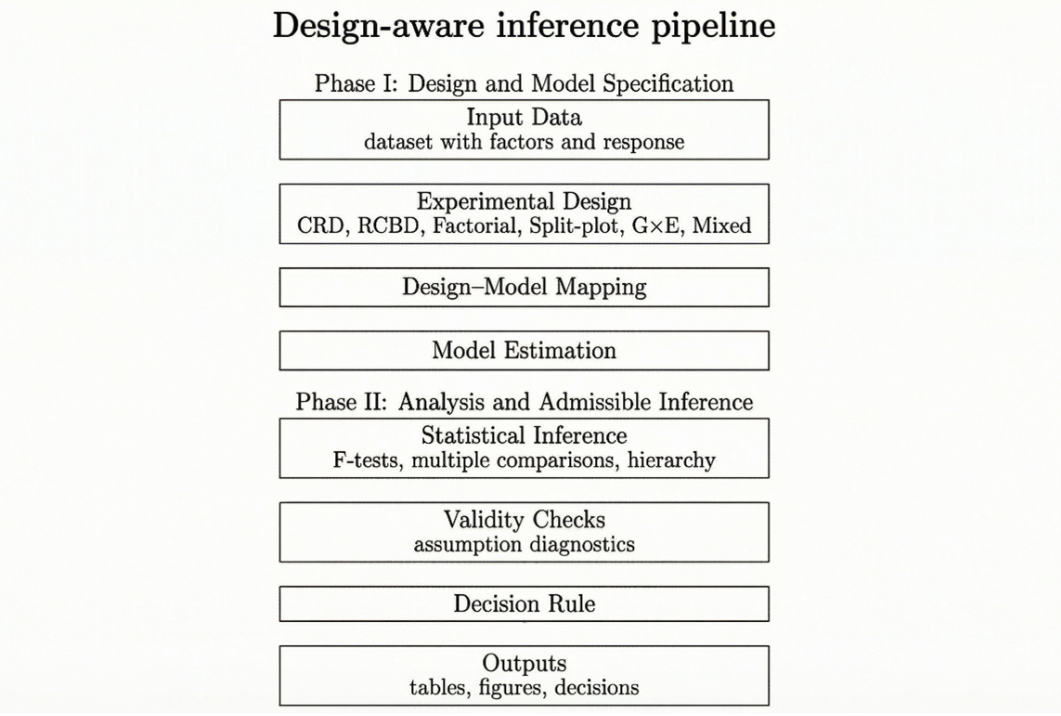

图1：在 AgroDesign 中实现的设计感知推断流程。第一阶段将实验设计映射到适当的统计模型。第二阶段执行允许的推断、假设验证和决策生成。

在规范阶段，会说明实验设计，并将其映射到相应的统计模型。系统根据实验设计、因子层级和实验单位构建正确的模型结构，并使用经过验证的

数值算法进行参数估计。这一步会识别可估效应，但暂不下结论。在推断阶段，会使用与各因素相对应的适当误差层来评估统计检验。多重比较程序和分层解释规则会限定有效的比较范围。随后，通过诊断检验检查模型有效性，只有统计上有效的结论才会提交给决策规则。最终结果是一组基于约束推断空间而非统计表格得出的结构化结论。因此，实验设计充当了分析的可执行规范。实现的设计谨慎地区分了统计方法学与数值计算。参数估计基于经过验证的统计算法，而框架则负责管理模型结构、推断有效性和解释逻辑。这种区分既保证了数值精度，也确保统计推断与实验设计保持一致。

4 实验验证

所有验证示例均表明，结果与经典实验设计理论一致。为确保所提出的框架得出的推断在统计上有效，而非仅仅复现软件结果，我们对一组代表不同实验结构的受控农业实验进行了分析。每个数据集都代表一种农业研究中常用的典型设计类别：完全随机实验、区组试验、析因处理、分层田间设计以及多环境育种试验。验证聚焦于以下四项标准：

1. 正确识别统计模型和分母误差。
2. 正确限定解释域。
3. 均值分离结果一致。
4. 基于允许的推断得出合理的农艺建议。

本框架的目标并非评估预测精度，而是确保方法正确性：应用经典实验设计理论时，框架应得出与训练有素的统计学家相同的结论。

4.1 完全随机设计（CRD）

第一项验证涉及一个完全随机实验，其中处理独立地施用于相似的实验单位。在该实验中，观测值由模型（1）描述，其中处理效应与共同的残差方差进行比较。该模型会自动构建固定效应模型，并生成表1所示的方差分析表。处理效应高度显著（F值 $= 145.33$，p值 ¡ 0.001），这证实了其存在

表1：为 CRD 实验自动生成的方差分析表。

来源	DF	MS	F	p值
处理	3	363.333	145.333	< 0.001
残差	16	2.500	–	–

肥料处理之间的产量差异。由于该设计只包含一个因素，且没有任何层级结构，可接受的解释域仅包括处理主

效应。基于图基 HSD 检验的均值比较表明，存在四个不同的组，且 NPK100 产生的平均产量最高。由于不存在交互结构，处理比较不受任何限制。处理比较见图 2。

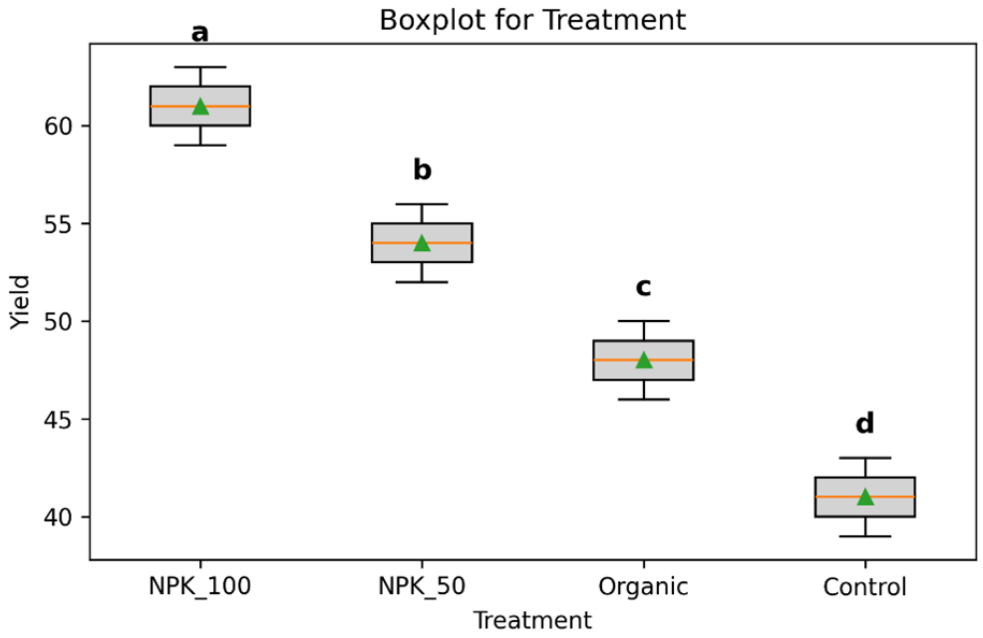

图 2：随机区组设计实验的品种比较与紧凑字母显示。字母表示在调整区组效应后、于 $\alpha = 0.05$ 水平下的图基 HSD 检验分组。

该图验证了均值分离检验的结果。字母不同的处理具有互不重叠的置信区间，且中位数的排序与方差分析检验结果相符。该图表明，该框架能够直接从统计模型的决策层生成达到发表质量的农学分析。残差图会自动生成。夏皮罗-威尔克检验显示正态性存在轻微违背（p值 $= 0.034$），说明该框架能够正确指出假设违背，而不改变模型结构或决策逻辑。

4.2 Blocking验证（随机完全区组设计（RCBD））

为检验该框架是否能恰当地处理区组结构问题，我们使用了一个随机完全区组设计（RCBD）数据集。在随机完全区组设计中，比较处理时需要先剔除区组间的变异，因此，如表 2 所示，处理方差与区组内的残差方差进行比较。处理效应极显著（$F = 320.00$，$p < 0.001$），区组效应也显著，证实了区组间存在环境异质性。均值分离识别出四个不同的品种组，其中品种 V4 的产量最高。品种分离情况如图 3 所示。

表 2：自动生成的随机区组设计实验方差分析。

来源	DF	MS	F	p值
品种	3	106.667	320.000	< 0.001
区组	3	5.667	17.000	< 0.001
残差	9	0.333	–	–

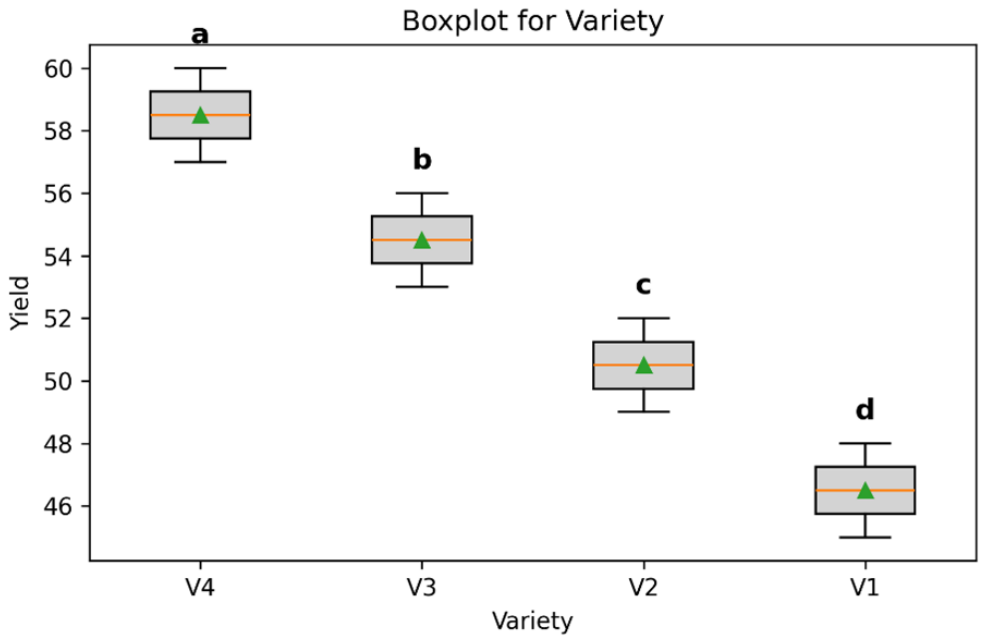

图 3：随机区组设计实验中经区组调整后的品种比较。字母表示在 $\alpha = 0.05$ 下的图基 HSD 检验分组。均值差异反映区组内对比。

调整区组变异后，各品种间的差异仍然清晰可辨。图形展示基于恰当误差项进行的处理比较，表明在比较各处理之前已将干扰变异纳入考量。由于该实验设计中不存在交互结构，因此在调整区组变异后，分析止于主效应层面。残差图显示数据没有偏离正态分布（夏皮罗-威尔克检验 p值 $= 0.064$），这证明模型拟合有效。

4.3 析因实验（交互作用感知推断）

为检验多因素模型的自动构建，考察了一个包含氮和植物间距的双因素析因数据集。在析因试验中，同时考虑主效应和交互作用效应，结果基于最高阶显著效应。该析因实验的方差分析结果见表3。氮和间距的主效应均高度显著，而互作项不显著（F $= 1.50$，p $= 0.262$）。由于交互作用在统计上不显著，可接受的解释域由两个主效应构成。响应趋势的相似性表明不存在交互作用（图4，方差分析结果也证实边际解释是有效的。

表3：为析因实验自动生成的方差分析。

来源	DF	MS	F	p值
氮	2	433.500	433.500	< 0.001
间距	1	112.500	112.500	< 0.001
氮 × 间距	2	1.500	1.500	0.262
残差	12	1.000	–	–

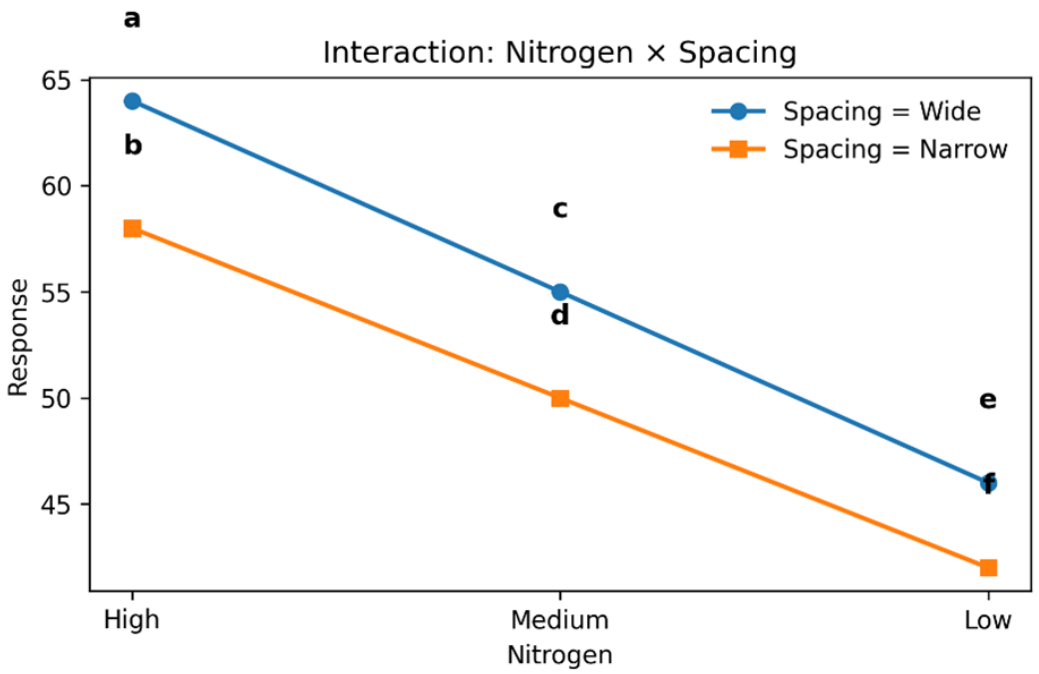

图 4：析因实验中氮 × 间距的交互作用图。响应曲线几乎平行，表明不存在交互作用，因此支持对主效应进行边际解释。

因此，该框架允许进行独立的主效应解释。均值分离表明氮有三个不同水平，间距有两个水平。由于没有占主导地位的交互作用，处理的排序采用边际均值。最高产量来自处理“高 × 宽”，但由于不存在显著交互作用，该结果可基于独立的主效应解释加以论证。残差分析表明偏离正态性（夏皮罗-威尔克p值 $= 0.0016$）。

4.4 裂区试验（多层推断验证 n)

裂区设计包含多个实验单位，因此不同因子需要使用不同的分母均方。主区因子必须以主区误差项进行检验，而副区因子及其交互作用则以副区残差误差进行检验。误判这些层是应用研究中推断错误的常见来源。裂区试验的方差分析结果见表4。灌溉与品种之间的交互作用具有统计显著性，而主区误差项（区组 × 灌溉）则不显著。灌溉 × 品种交互作用高度显著（$F = 12.0$，$p < 0.001$），这表明

表4：自动生成的裂区试验方差分析。

来源	DF	MS	F	p值
区组	2	8.111	73.0	< 0.001
灌溉	2	626.333	5637.0	< 0.001
品种	2	90.333	813.0	< 0.001
区组 × 灌溉	4	0.111	1.0	0.445
灌溉 × 品种	4	1.333	12.0	< 0.001
残差	12	0.111	–	–

品种表现取决于灌溉水平。因此，解释仅限于处理组合，而非边际主效应。这证实该框架能够在确定可接受的解释域时，正确优先考虑最高阶显著效应。图5展示了交互结构。响应线并不平行，这证实品种排序会随灌溉水平变化，支持交互作用主导的推断。

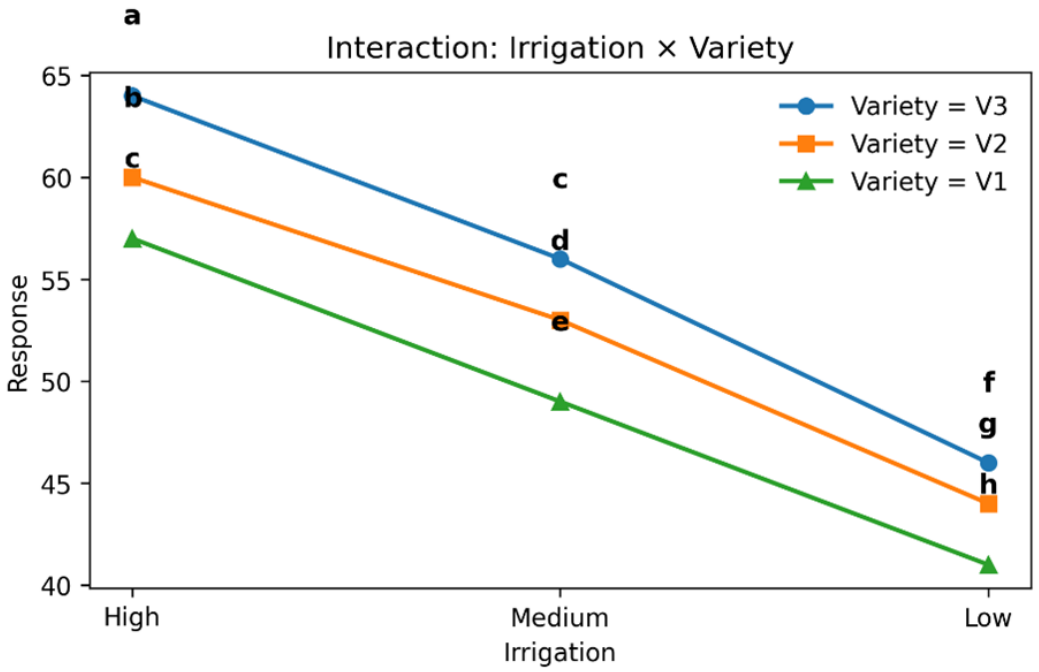

图5：裂区试验中的灌溉 ×品种交互作用。响应曲线不平行表明应采用交互作用主导的推断，并在组合层级上进行解释。

大部分变异来自随机区组效应，因此有理由采用经方差调整的处理排序。这说明了从固定效应比较到方差分量推断的自动过渡。在对区组变异进行调整后，处理T3具有最大的调整后表现。由于处理排序是基于预测值而非边际均值进行的，因此建议采用经方差调整的处理效应。

4.5 线性混合模型验证

一些农业试验需要将环境或区组效应建模为随机效应而非固定效应。为检验方差分量估计和处理排序，对一个混合效应数据集进行了分析，其中区组作为随机效应，处理作为固定效应。自动估计的方差分量见表5。正的区组方差证实

表5：线性混合模型估计的方差分量。

分量	方差
区组（随机）	1.22
残差（随机）	0.44

环境异质性的存在为采用混合模型设定提供了依据。我们使用最佳线性无偏预测值（BLUPs）评估处理表现，这些预测值会根据随机区组变异对处理均值进行调整。基于最佳线性无偏预测值的处理表现汇总于表6。图6展示了区组方差分量和残差方差分量的相对贡献。

表6：基于最佳线性无偏预测的混合模型处理排序。

处理的最佳线性无偏预测值	
T3	54.5
T2	48.5
T1	41.5

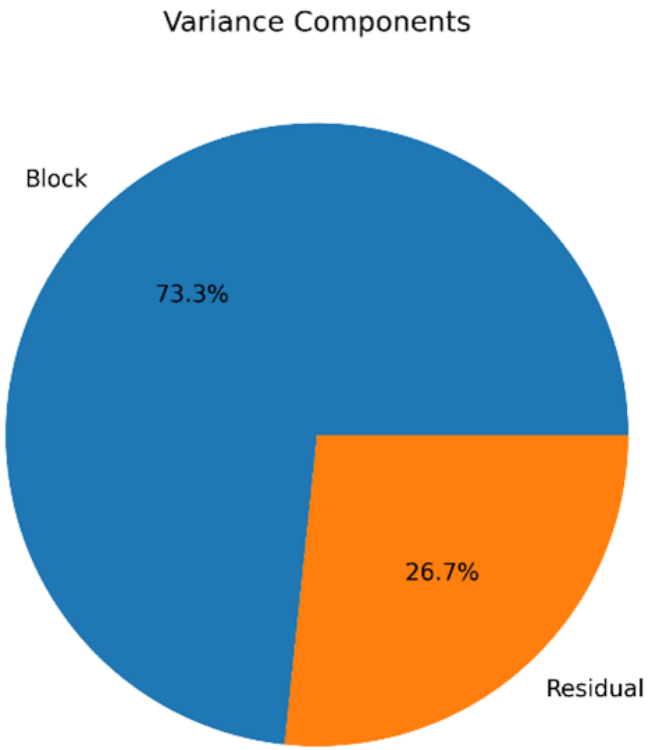

图6：线性混合模型下总方差中可归因于区组与残差成分的比例。

混合模型的结果表明，处理的排序依据的是方差估计值，而非原始边际均值。将区组效应作为随机效应纳入模型后，推断方式从固定效应方差分析转变为最佳线性无偏预测推断，同时界面保持不变。这一结果符合预期，即该系统将传统的方差分量分析与设计指导的决策逻辑相结合。在统一设计规范下能够切换固定效应推断和混合效应推断，凸显了该框架的灵活性。

4.6 多环境试验（G×E）验证

为评估跨环境表现，分析了一个基因型×环境（G×E）数据集。在此类试验中，结论取决于互作项的大小：强交互作用意味着需要环境特异性建议，而弱交互作用则允许总体排序。多环境试验的方差分析结果见表7。基因型 ×环境效应显著，而互作项不显著（$p = 0.535$）。因此，处理排序在各环境间均有效，并可解释基因型边际表现。框架据此进行基于最佳线性无偏预测值的排序（表8）：

图7以图形方式展示了基于最佳线性无偏预测的基因型表现。

表7：自动生成的多环境试验方差分析。

来源	DF	MS	F	p值
基因型	3	364.458	728.917	< 0.001
环境	3	42.458	84.917	< 0.001
基因型 $\times$ 环境	9	0.458	0.917	0.535
残差	16	0.500	–	–

表8：跨环境基于基因型最佳线性无偏预测值的排序。

基因型最佳线性无偏预测
G4 7.7
G3 2.7
G2 -2.3
G1 -8.1

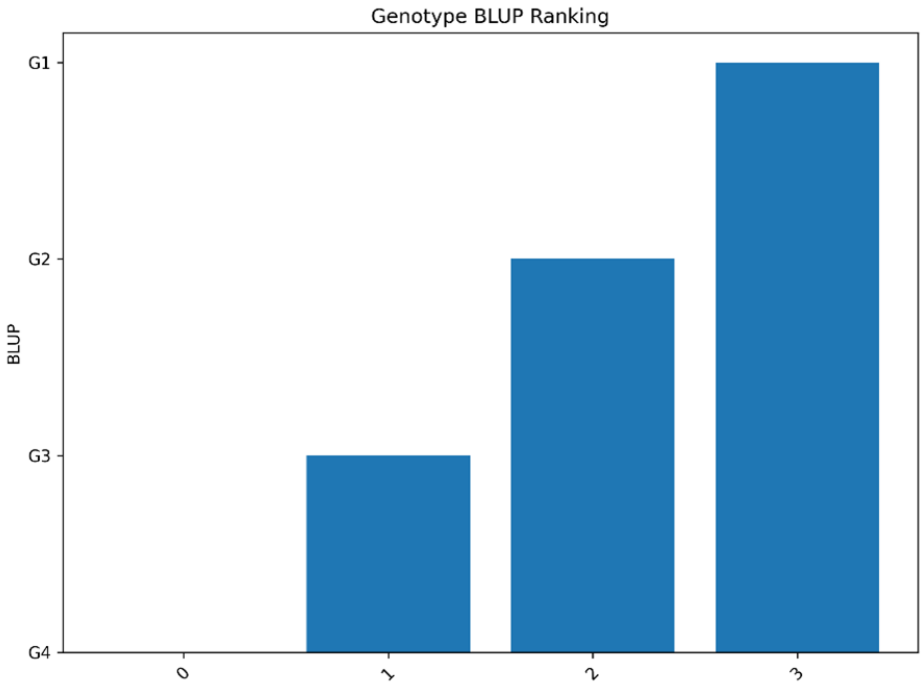

图7：跨环境基于最佳线性无偏预测值的基因型排序。正值表示表现高于平均水平。

跨环境的预测表现显示，基因型G4表现优异，这与交互作用不显著的预期一致。基因型G4的预测表现最高。广义遗传力估计为 $H^2 = 0.99$，这反映出遗传控制较强，且选择有效。由于交互作用不显著，稳定性分析证实了其广泛适应性，而非在个别环境中的优势。双标图（图8）显示存在单一超级环境，且基因型表现稳定；这与广泛适应性的统计结果一致，也为从方差分析中提取的决策规则提供了依据。

结构。因此，该模型提出的是一个全局基因型，而非针对特定环境的推荐方案。残差分析表明数据符合正态性（$p = 0.121$），这证明了模型的有效性。

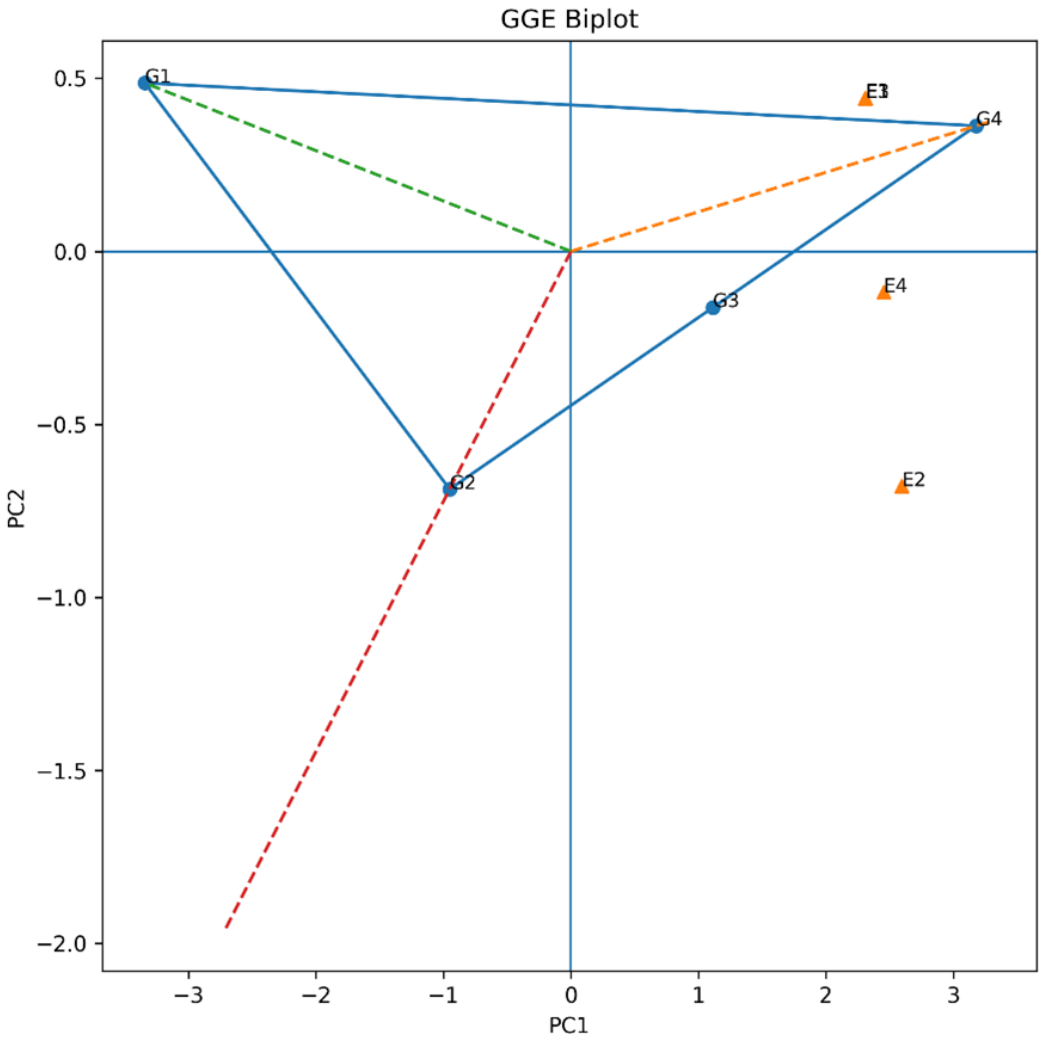

图8：基因型-环境表现的GGE双标图。与平均环境轴的距离较近表明具有广泛适应性。

5 讨论

本文阐明了实验设计的一项基本原则：统计推断与实验随机化框架之间的关系 [2, 6]。经典理论确立了这种关系，但当代计算分析往往需要通过回归分析手动纳入这种关系。AgroDesign 将设计语义融入分析工作流，弥合了这一差距。其中一个最重要的方面是强制执行分层解释规则。在析因设计和裂区设计中，存在高阶交互作用效应时，低阶效应可能无法解释 [5]。在标准统计软件中，所有效应都会同时计算并报告。当前系统

将这条规则形式化为确定性计算约束：允许的解释范围会自动缩小至最高阶显著效应。因此，统计结果在结构上变得可解释，而非仅仅停留于数值描述。该系统还在一个设计感知框架中结合了固定效应推断与混合模型推断。农业试验常涉及环境条件的变异，因此需要进行方差分量估计和预测，而非均值比较。使用当前系统时，可以通过同一界面将因子指定为随机因子，工作流随后会自动从基于方差分析的推断切换为基于最佳线性无偏预测的估计，如混合模型分析所示 [7]。这使分析师在从田间试验转向育种试验时，无须改变分析范式或更换软件平台。多环境分析展示了另一项优势：将统计交互作用结构直接映射到决策逻辑。是否存在基因型与环境的交互作用，决定了全局建议还是环境特定建议才有效。加性主效应与乘性交互作用模型等稳定性方法以及基于回归的方法通常作为独立分析开展。在当前系统中，它们成为同一结构模型的分析结果，从而确保稳定性分析与处理排序在逻辑上保持一致。从软件开发的角度来看，主要创新并非提供新的统计估计量，而是将实验设计编码为可执行规范。研究者不再根据对数据的检查来选择检验，而是指定实验设计，之后可能进行的统计操作便会确定性地随之而定。这提高了规范唯一性，尤其是在层级试验中；在这类试验里，错误选择分母仍是应用实践中的常见错误。该系统还促进计算可重复性。分析的灵活性和隐藏的建模假设被认为是造成不可重复科学结果的原因 [11]。通过对模型构建、解释和比较施加受规则约束的工作流，该系统消除了分析师特有的变异，同时保持与通用 Python 数据科学工作流的兼容性。不过，系统仍存在若干局限。目前的验证仅限于平衡典型设计，未来应扩展至非平衡设计和缺失数据田间试验。目前的假设检验仅涉及正态性和方差齐性。未来研究可将该框架推广至广义线性模型和贝叶斯模型。最后，自动化解释并不能取代农学知识，而是为专家知识发挥作用提供统计上一致的基础。总之，本研究表明，将实验设计表示为计算对象，是使统计理论与软件开发相协调的一种方式。本文开发的系统要求进行恰当的推断，将经典方差分析与混合模型相结合，并在当代计算环境中提供可重复的农学建议。

6 结论

本研究提出了一种用于统计推断的设计驱动框架，其中实验设计是分析的核心规范。系统不会要求分析师手动指定统计模型，而是仅根据实验中声明的随机化结构，自动推断模型、假设检验和解释规则所允许的结构。该框架将固定效应方差分析、层次设计、混合模型和多环境稳定性分析整合到一个声明式分析流水线中。通过在推断过程中纳入层级解释约束并正确识别误差层，系统确保所有结论都符合已知的实验设计原理。因此，该软件将统计分析从一系列由分析师作出的选择，转变为从实验结构到允许的推断和

建议的确定性转换。针对传统农业试验设计进行的实验验证表明，该系统能够正确复现经典统计分析所得出的结论，同时消除规范指定上的歧义。系统输出的处理比较和农艺建议等内容，依据的是统计上有效的比较范围，而非模型输出。将设计语义纳入可执行计算，能够增强可复现性并减少变异。未来的开发将扩展对广义响应分布、非平衡田间试验和更大规模多地点育种数据集的支持。尽管如此，当前框架为在 Python 中开展可复现、设计感知的统计分析奠定了基础，使实验结论能够直接源自所实施的实验。

7 软件可用性与可重复性

AgroDesign 以开源 Python 软件包的形式分发，并通过 Python 包索引（PyPI）公开提供。可使用以下命令安装：

```
pip install agrodesign
```

完整的源代码、文档、版本历史记录和问题跟踪器均维护在公开的 GitHub 代码仓库中：

代码仓库：
https://github.com/DeepStatistix/AgroDesign

实验验证中使用的所有数据集均包含在软件包内，作为内置教程数据集，代表典型的实验设计，包括完全随机设计（CRD）、随机完全区组设计（RCBD）、析因试验、裂区设计、线性混合模型和多环境试验（G×E）。这些数据集均为合成数据，仅用于方法验证和可重复性研究。

本文展示的每项分析均可通过软件包内置的数据集加载器和默认工作流程以编程方式复现。该框架会自动生成方差分析表、方差分量估计、多重比较结果、诊断检验、图形可视化和处理排序汇总。

由于模型设定直接源自已声明的实验设计，分析流程具有确定性。因此，本文报告的所有统计结果都与公开代码库中的可执行代码完全对应。本文阐述了该框架所依据的方法论和结构原则，而代码库则提供了完整的计算可重复性。

参考文献

[1] de Mendiburu, F.（2021）。agricolae：农业研究统计方法（版本 1.3-5）。拉莫利纳国立农业大学，利马，秘鲁。[2] Gomez, K. A., & Gomez, A. A.（1984）。农业研究统计方法（第2版）。约翰·威立父子公司。[3] Garijo, D., Alper, P., Belhajjame, K., Corcho, O., Gil, Y., & Goble, C.（2014）。科学工作流中的常见模式：一项实证分析。《未来一代计算机系统》，36，338–351。[4] Morris, M.（2010）。实验设计：基于线性模型的导论。Chapman & Hall/CRC。[5] Montgomery, D. C.（2017）。实验设计与分析（第9版）。约翰·威立父子公司。[6] Steel, R. G. D., Torrie, J. H., & Dickey, D. A.（1997年）。统计学原理与方法：生物统计学方法（第3版）。麦格劳-希尔。[7] Piepho, H.-P., B¨uchse, A., & Emrich, K.（2003）。农业试验中非平衡混合模型分析的统一方法。《农学与作物科学杂志》，189(4)，209–218。[8] Seabold, S., & Perktold, J.（2010）。Statsmodels：使用 Python 进行计量经济学与统计建模。第*9*届 *Python*科学会议论文集，92–96。[9] Virtanen, P., 等（2020）。SciPy 1.0：Python科学计算的基础算法。《自然方法》，17，261–272。[10] Terpilowski, M.（2019年）。scikit-posthocs：Python 中的两两多重比较检验。《开放源代码软件杂志》，4(36)，1169。[11] Stodden, V., Leisch, F., & Peng, R. D.（2013）。实现可重复研究。《科学与工程计算》，15(4)，8–12。[12] Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E.（2013）。可重复计算研究的十条简单规则。《*PLoS*计算生物学》，9(10)，e1003285。

最小可复现示例

本附录演示了如何使用 AgroDesign 框架完成一个完整的分析工作流。此示例展示了如何通过单一声明式规范，将随机完全区组设计（RCBD）试验从原始数据分析到农艺建议。

安装

该软件包可从 Python 包索引（PyPI）公开获取，并可使用以下命令安装：

```
pip install agrodesign
```

加载内置数据集

AgroDesign 提供了用于表示典型农业试验设计的内置数据集。以下示例采用随机完全区组设计进行品种试验。

```
from agrodesign.datasets import load_数据集
from agrodesign.experiment import Experiment

df = 加载_数据集("rcbd")
df.head()
```

该数据集包含多个作物品种在不同区组中评估所得的产量观测，这些区组代表了田间异质性。

运行分析

用户指定的是实验结构，而不是统计模型：

```
结果 = Experiment(df, "Yield").rcbd("Variety", "Block").run()
```

该框架会根据所声明的设计自动确定合适的线性模型，构建方差分析表，并应用正确的误差项。

统计报告

完整的统计报告可以直接打印：

```
print(result)
```

此命令会自动生成：

- 方差分析表
- 多重比较结果
- 解释层级
- 假设诊断

农艺建议

该　可以独立于统计报告进行农学解读　t:

result.summary()

这会生成一份简洁的面向决策的摘要，且符合可接受的解释域。

可视化

自动生成可供发表的图表：

result.plot()

导出可复现输出

所有表格和图表均可导出到结构化结果目录：

result.export("results")

导出的文件夹包含：

- 格式化方差分析表
- 处理比较摘要
- 诊断图
- 完整文本报告