**对*Constant (2023)*研究结果的计算可复现性检验**

小组成员分工

|  |  |  |  |
| --- | --- | --- | --- |
| 组长 | 邬思宇 | | |
| 组员 | 邬思宇、师淑婷、程蕾 | | |
| 分工 | | | |
| 数据分析 | 邬思宇、师淑婷、程蕾 | PPT 制作 | 邬思宇 |
| 文字报告制作 | 邬思宇、师淑婷、程蕾 | PPT 展示 | 邬思宇 |

\* 同一名同学可负责多个部分；如同一内容由多位同学负责，可按百分比注明贡献占比

**摘要**：根据贝叶斯模型，决策与信心评估均基于先验期望与即时信息经过精度加权后的整合结果。该理论假设先验信息在决策与信心评估中得到同等且最优整合，但此假设尚未被验证。研究通过三项实验量化了先验信息对决策与信心的影响机制：采用双决策任务创造后验信息相同但先验/即时信息权重不同的条件对。结果发现，在辨别决策中先验信息被低估，但在对应决策的信心评估中低估程度减弱，且这种现象与处理时间无关。当采用外源性概率线索作为先验时，该模式依然存在。研究通过贝叶斯模型量化两个层面的先验权重参数，获得一致性证据：尽管先验信息在决策中未被充分利用，但在显性信心评估中仍能得到更优化的运用。

**关键词**：贝叶斯推理；先验信念；信心评估；计算可复现性；

**1 引言**

**1.1 所选文献信息**

**表 1 文献信息表**

|  |  |  |  |
| --- | --- | --- | --- |
| **1 文献基本信息** | | | |
| 所选文献 | Constant, M., et al. (2023). Prior information differentially affects discrimination decisions and subjective confidence reports. *Nature Communications*, *14*(1), 5473. https://doi.org/10.1038/s41467-023-41112-0 | | |
| 数据来源 | https://gitlab.com/MarikaConstant/priors-in-confidence | | |
| **2 文献选取** | | | |
| 文献主题是否包含不止一篇研究？ | * 是，且包含元分析研究 * 是，但不包含元分析研究 * 否 | 文献此前被其他研究者重复过？ | * 是(附上原文链接) * 否 |
| **3 研究假设选取** | | | |
| 重复的研究假设 | 假设一：如果先验被最优整合，目标正确率主要受后验信息水平影响:如果先验被低估，正确率主要受实验条件影响；  假设二：如果先验未被最优整合，不同条件类型对正确和错误试次的影响将会不同。 | | |
| 重复的研究假设是否在其他研究中经过重复？ | * 是(附上原文链接) * 否 | 文献共几个实验，重复的研究假设是第几个实验中的？ | 共三个实验，复现的研究假设均来自实验一 |
| 选择该假设的原因 | 是本研究的核心假设，也是整个实验的基础 | | |
| **4 数据集选取** | | | |
| 是否采用原始数据？ | * 是 * 否 | 是否对样本量进行修改？ | * 是(说明原因) * 否 |
| 若修改，报告原文样本量大小和修改后的样本量大小 |  | 若修改，报告使用 G-power 计算的修改后的样本量对应的效应量 |  |

**1.2 文献介绍**

**1.2.1 研究背景**

人类的感知和决策常被描述为贝叶斯推理过程。从计算角度而言，这一过程涉及将先验分布与似然分布进行整合，进而得到后验分布，该分布随后构成信念或感知的基础。并且，信心，即通常伴随感知决策而产生的确定性感受，同样可通过贝叶斯模型（Bayesian inference models）来进行解释。

传统贝叶斯模型认为，人类的感知和决策是基于先验期望与新信息的精度加权整合。在这一框架下，同一后验分布同时驱动决策与信心。其依赖于两个关键假设：（1）先验和新信息在决策和信心中均被最优地整合；（2）二者的整合机制相同。

然而，这些假设尚未经过实证检验，尤其是在信息性先验（informative priors）条件下，先验的权重是否存在偏差，以及决策与信心是否存在整合差异仍不明确。已有研究发现，信心可能整合了与决策不同的信息（如决策后的证据积累），且呈现“folded-X”模式，即信心与决策正确性的交互作用。这提示决策与信心在信息处理上可能存在不对称性。但这种不对称性是否延伸至先验信息的整合过程，仍缺乏直接证据。此外，以往研究多使用平坦先验（flat priors，即无信息先验），无法检测先验整合的潜在偏差；近期研究虽引入信息性先验，但仅发现其影响信心阈值，未量化其权重或比较决策与信心的差异。

**1.2.2 研究问题**

基于以上原因，该研究通过三个行为实验和计算建模，旨在探讨：（1）在先验信息性较强的情况下，人们是否能最优地整合先验与新信息？若不能，是表现为对先验的权重不足还是权重过度？（2）先验在决策和信心中的权重是否存在差异？

**1.2.3 研究假设**

*H*1：如果先验被最优整合，目标正确率主要受后验信息水平影响；如果先验被低估，正确率主要受实验条件影响；

*H*2：如果先验未被最优整合，那么在目标刺激的正确和错误试次下，实验条件会以不同方式调节后验信息对信心的影响。

**1.2.4 研究结果**

先验信息在决策中的权重较低，而在信心中的权重相对较高，更接近最优整合状态。

**2 方法**

**2.1 样本**

实验一共有21名被试的数据被纳入分析，其中包含10名男性和11名女性，年龄在18至37岁之间（*M* = 25.7，*SD* = 4.6）。所有被试均为右利手（爱丁堡惯用手量表评分：*M* = 83.2，*SD* = 28.5）。

实验二共有25名被试的数据被纳入分析，其中包含9名男性，15名女性，以及1名未说明性别的被试。被试年龄在19至34岁之间（*M* = 25.4，*SD* = 3.8）。所有被试均为右利手（爱丁堡惯用手量表评分：*M* = 86.3，*SD* = 34.3）。

实验三共有20名被试的数据被纳入分析，年龄在19至33岁之间（*M* = 23.7，*SD* = 3.9）。所有被试均为右利手（爱丁堡惯用手量表评分：*M* = 85.4，*SD* = 39.0）。

所有被试视力或矫正视力正常，英文流畅，在开始实验前签署过知情同意书。

**2.2 原研究方法简介**

**2.2.1 研究设计**

实验一和实验二为3×2被试内设计。第一个自变量为后验信息总量，包含低、中、高三个水平；第二个自变量为条件类型，包含强先验和强似然两个水平。因变量为第二次决策，即目标决策的正确率和信心评分。

在这两个实验中，每个trial包含两次连续的随机点运动刺激，被试需判断这两次随机点运动刺激的运动方向（左/右），并对第二次判断进行信心评分，信心程度为50%-100%，50%代表猜测，100%代表完全确定。实验二中的范式保持不变，但在进行目标决策之前增加了2秒的延迟期。在这2秒期间，被试被指示尽量避免按键，并在每个block结束时收到过早按键次数的反馈。被试一共需要完成两个session，每个session下10个block，每个block包含36个trial。在每个session开始前，还需完成一个控制任务，包含90个trial，控制任务只呈现一次点刺激并进行判断。其目的是用于估计个体内部噪声和决策偏差。

实验三为单因素被试内设计，自变量为先验概率水平，包含60%，70%，80%和90%四个水平。因变量为决策正确率和信心评分。在实验三中，被试只需进行一次随机点运动方向的判断，在每次刺激前被明确告知刺激运动方向的概率。每种概率条件下100个trial，在10个block间交错平衡。此外还有100个trial为无信息性先验，概率为0.5，用于估计个体内部噪声和决策偏差。

**2.2.2 数据分析**

同时采用频率主义和贝叶斯主义的框架进行假设检验。使用R语言中lme4包构建混合效应模型。模型包含固定效应和随机效应，通过比较模型与零模型得到*χ*²值和*F*值。同时，还使用BayesTestR和brms包计算贝叶斯因子（*BF*10）。

为更直接和定量地比较先验在这两个过程中的差异，研究通过构建并拟合贝叶斯生成模型量化两个独立参数，和。模型通过Stan和cmdstanr包建构并编译，并通过MCMC方法进行拟合。

**2.3 重复思路说明**

实验一通过双决策任务构建了先验与似然信息对比的条件，是验证研究核心问题的基础，其结果直接支撑文章核心结论。因此对实验一进行重复性检验，以验证结果可靠性。

实验一数据分析需复现行为统计分析与计算模型拟合两部分，以验证“先验在决策中被低估，在信心被更优使用”的核心结论。使用原文献方法进行重复。根据作者提供的脚本说明文件，按顺序先进行行为分析，再估计单个被试的内部噪声与决策偏差并添加数据列，通过构建分层模型拟合数据得到具体的参数值并绘制后验分布。由于使用的Stan版本不同，源代码中Stan模型使用的是旧的数组声明语法，因此需要将旧语法改为新语法。

**3 结果**

**3.1 描述性统计**

实验一描述性统计结果的比较如表2所示：

**表 2 实验一描述性统计结果的比较（*N*=21）**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 变量（目标正确率） | 原研究报告结果 | | 本研究 | | δ | 评级 |
|  | Mean | *SD* | Mean | *SD* |
| L\_Stronger\_Target | 0.72 | 0.05 | 0.72 | 0.05 | 0% | 完全一致 |
| M\_Stronger\_Target | 0.79 | 0.04 | 0.79 | 0.04 | 0% | 完全一致 |
| H\_Stronger\_Target | 0.84 | 0.04 | 0.83 | 0.04 | 1.20% | 次要偏差 |
| L\_Stronger\_Lead | 0.70 | 0.05 | 0.70 | 0.05 | 0% | 完全一致 |
| M\_Stronger\_Lead | 0.76 | 0.05 | 0.76 | 0.05 | 0% | 完全一致 |
| H\_Stronger\_Lead | 0.81 | 0.05 | 0.81 | 0.05 | 0% | 完全一致 |

**3.2 推断性统计**

**3.2.1 使用与原文献相同方法的推断性统计**

推断性统计结果的比较如下表3和表4所示：

**表 3 假设一推断性统计结果的比较**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 后验信息水平主效应 | | | 条件类型主效应 | | |
|  | 统计量  (*χ2*) | 显著性指标(*p*) | 其它指标(*BF*10) | 统计量  (*χ2*) | 显著性指标(*p*) | 其它指标(*BF*10) |
| 原文献  报告结果 | 175.77 | < 0.001 | 6.12×1036 | 9.06 | 0.003 | 8.95 |
| 本研究 | 175.77 | < 0.001 | 6.02×1036 | 9.06 | 0.003 | 9.03 |
| *δ* | 0% | 0% | 1.63% | 0% | 0% | 0.89% |
| 评级 | 完全一致 | 完全一致 | 次要偏差 | 完全一致 | 完全一致 | 次要偏差 |

**表 4 假设二推断性统计结果的比较**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 后验水平和正确性的交互作用 | | | 后验水平、正确性与条件类型的交互作用 | | |
|  | 统计量  (*F*) | 显著性指标(*p*) | 其它指标(*BF*10) | 统计量  (*F)* | 显著性指标(*p*) | 其它指标(*BF*10) |
| 原文献  报告结果 | 24.5 | < 0.001 | 2.84×106 | 2.66 | 0.07 | 70.83 |
| 本研究 | 23.53 | < 0.001 | 5.53×1010 | 2.66 | 0.07 | 120.01 |
| *δ* | 3.96% | 0% | 1.947×106% | 0% | 0% | 69.43% |
| 评级 | 次要偏差 | 完全一致 | 主要偏差 | 完全一致 | 完全一致 | 主要偏差 |

模型拟合结果的比较如下表5所示

**表 5 模型拟合的组水平后验均值结果比较**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 样本量(*N)\** |  |  | *b* |
| 原文献  报告结果 | 20 | 2.17 | 1.27 | 2.18 |
| 本研究 | 20 | 2.17 | 1.26 | 2.18 |
| *δ* | 0% | 0% | 0.79% | 0% |
| 评级 | 完全一致 | 完全一致 | 次要偏差 | 完全一致 |

\* 原本有21名被试，其中1名因参数异常被排除。作者先对单个被试进行拟合，其中1名被试（其索引为9）的参数与组均值的偏差达到6.05个标准差，故排除。

**3.3 对原文计算可复现性进行评估**

**3.3.1 使用与原文献相同方法**

报告原文献的值的评级分布、推论的一致情况，整理成表格，如下表所示：

**表 5 结果可复现性的评估表**

|  |  |  |
| --- | --- | --- |
| 可复现性情况 | 数量及占比 | |
| *N* | *%* |
| 完全一致(*δ* = 0%) | 15 | 68% |
| 偏差较小(0% < *δ* < 10%) | 5 | 23% |
| 偏差较大(*δ* ＞ 10%) | 2 | 9% |
| 因舍入导致的偏差 | 2 | 9% |

\* 结果数量*N*指在重复分析中，对重复分析结果与原结果进行配对比较的次数。对于每个目标效应，结果包括一组数值，如汇总效应估计(summary estimate，如*t*值/ F值)、置信区间界限(confidence interval bound)、效应量(effect size)样本大小(size effect)等，应将原文中报告的每个数值与重复结果进行比较。例如，在一个*t*检验中，原文献报告了*t*值、95%置信区间、cohen’s d和样本大小，则这个效应中*N*＝4。将各效应的*N*求和即为全体数量。

**表 6 推论的一致性的评估表**

|  |  |  |
| --- | --- | --- |
| 推论的一致性 | 数量及占比 | |
| *N \** | *%* |
| 一致 | 11 | 100% |
| 不一致 | 0 | 0% |

\* 推论数量*N*指在重复分析中，对效应做出统计推断的次数。例如，仅进行了一个*t*检验，则*N*＝1；如果进行了一个2\*2的方差分析，并进行简单效应分析，则有可能有7个统计推断：两个主效应的推论，一个交互作用的推论，四个可能的简单效应分析的推论，因此*N*＝7。如果报告的*p*值相对于重复的*p*值落在显著性水平边界的另一侧，则被归类为推论不一致；反之为推论一致。

**4 讨论**

**4.1 计算可复现性检验结果分析**

结合下表，对原文献进行分析，推测可能导致可复现性检验结果差异的原因。对于重要的原因，逐段进行展开说明。

**表 8 计算上（不）可重复的原因分析表**

|  |  |  |  |
| --- | --- | --- | --- |
| **可能原因** | | | **研究一** |
| **原文献开放性问题** | **一般性开放获取问题** | 几个结果的微小差异，可能是由于分析中使用了没有设置固定种子的随机数； | 源代码中没有设随机种子 |
| 个别结果的微小差异，可能是由于印刷或复制粘贴错误； | 否 |
| 文章文本中程序报告不明确，包括纳入亚组的标准、缺乏或不正确报告用于回归模型的变量、以及未报告的单侧分析； | 否 |
| 在文章的开放实践声明中对研究的模糊标记。 | 否 |
| **OSF 开放获取特定问题** | OSF 中缺乏对数据和/或代码内容进行说明的文档(readme文档)； | 否 |
| OSF 上的数据与代码文件不一致，如代码中对部分数据进行了操作，但这部分数据在数据文件中无对应； | 否 |
| OSF上的数据存储问题，包括文件损坏或无法下载。 | 否 |
| **数据开放获取特定问题** | 没有提供原始数据； | 有原始数据 |
| 没有提供处理后的数据； | 有处理后的数据 |
| 没有提供数据处理过程的描述或代码。 | 有具体的分析代码 |
| **代码开放获取特定问题** | 缺乏共享的分析代码或建模代码； | 否 |
| 软件包或软件版本的问题。 | 是 |
| **重复过程的原因** | **重复研究与院研究的区别** | 是否使用同样的数据集； | 是 |
| 是否使用同样的数据分析软件及软件包； | 是 |
| 是否使用同样的数据分析方法。 | 是 |
| **重复者相关因素** | 重复者此前是否有过 R 使用经验； | 是 |
| 重复者对关于 R 的知识或操作上存在漏洞，较难理解原文章中的部分操作(可做简单说明)。 | 否 |
| **其他影响因素** | **文献年份** | 文献发表年份是否较为久远，是否在开放科学运动之前； | 文献发表年份为2023年 |
| **文献质量** | 文献引用量大小； | 12 |
| 是否有其他研究支持本文献结果； | 否 |
| 是否有其他研究对本文献结果进行了重复，重复结果如何(可做简单说明)。 | 否 |

结合重复结果，差异主要为贝叶斯因子（*BF*10）、少数描述性统计的均值及信心先验权重（）。

原研究虽提供代码，但未包含详细的环境配置文件（如R包版本、Stan 版本号），不同版本可能会导致结果存在差异。此外，原代码未固定随机种子，使得MCMC采样的初始值和迭代路径具有随机性，进而对贝叶斯因子这类对采样过程敏感的指标产生较大影响。

尽管部分统计值存在不一致，但最终的推论是一致的。核心效应的显著性和方向未发生改变，且关键参数的估计值在实际意义上仍支持“先验在决策中被低估、在信心中更优整合”的结论。因此，统计量的数值差异并未影响研究的核心发现与理论推断。

**4.2 其他思考**

通过本学期的R语言课学习，我们对数据分析有了更深的理解。总结来说，有以下几点：

第一，分析数据前要明确自己的研究问题和研究假设，要清晰界定编写和运行代码是为了解决什么问题，亦或是验证某个具体假设，目标的明确性是后续工作开展的基础与方向指引。

第二，数据作为分析的核心要素，其来源的可靠性、完整性直接影响最终结果，需仔细排查数据中的缺失值、异常值，做好数据预处理工作，同时关注数据格式是否适配分析需求，必要时进行格式转换。

第三，在代码编写过程中，代码的可读性和可维护性不容忽视，规范的命名方式、详细的注释不仅有助于自身对代码逻辑的梳理，也便于他人理解和复现。通过对这篇研究进行可重复性检验，让我们深刻意识到详细的脚本说明、规范的代码是多么重要。今后我们在完成自己的项目时，也会按照这种方式，方便自己也方便他人。

第四，R语言只是一门工具，重要的还是我们对基本的统计知识的理解与掌握。在学习过程中，我们常常会看见别人怎么做就去模仿，有时候跑出结果甚至都看不懂什么意思。这让我们意识到，单纯机械地复制代码而缺乏对统计原理的认知，就如同“知其然而不知其所以然”，扎实的基本功也是很重要的。胡老师的课常看常新，当我们带着一定的理解再去看之前的视频时，才明白原来这里是这么回事，原来还可以有不同的方法。

**参考文献**

Constant, M., et al. (2023). Prior information differentially affects discrimination decisions and subjective confidence reports. *Nature Communications*, *14*(1), 5473. https://doi.org/10.1038/s41467-023-41112-0