**NNU-R语言课程**

**分析与计算可重复性检验指南**



**初版整理人：** 周方茹 吕冬煦

**初版审核人：** 武婷婷 温佳慧

**1 概述**

**1.1 可重复性及其重要性**

近年，人们对科学研究可重复性(reproducibility)的关注持续上升，许多科学发现很难或无法复制，导致了“可重复性危机”(replication crisis)的出现。可重复性指研究人员使用与原始研究者使用的相同的材料复制先前研究结果的能力(Goodman et al., 2016)。可重复性是科学的决定性特征，不可重复性(irreplication)可能会破坏数据的重复使用(Hardwicke et al., 2018)，使复制尝试变得复杂(Nuijten et al., 2018)，并对科学证据的来源和真实性造成不确定性，从而有可能破坏任何相关推论的可信度(LeBel et al., 2018)。

由于可重复性是一个笼统的术语，有研究者试图将其内容划分为更易于处理的术语(Goodman et al., 2016)。基于教学目标，**本课程主要关注分析可重复性**(analytical reproducibility)或**计算可重复性**(computational reproducibility) 。

* **分析可重复性**指提供数据文件及对数据分析说明的情况下，其他研究者可以重现原研究中所报告的最终结果，包括关键的定量的结果、图表和数值(Kitzes et al., 2017)；
* **计算可重复性**指的是原研究者不仅提供数据及对数据分析的说明，而且提供分析所使用的代码，其他研究者可使用**原代码**和**原数据**重现原研究中所报告的最终结果，包括关键的定量结果、图表和数值。

**1.2 分析/计算可重复性的检验**

数据、材料和分析代码的共享是提高可重复性的重要举措，可使其他研究者复制、检验与建构现有的研究(Kitzes et al., 2017)。为了更好地促进对已有的研究的可重复性检验，*Psychological Science* 等期刊鼓励作者在 OSF(osf.io) 等平台上公开数据与代码，这些文献都可以进行复制与检验。

本课程作业需要通过R语言复现真实研究中的数据分析，帮助同学们在掌握 R 语言的同时，进一步了解与思考数据分析和计算可重复性的问题。通过复现分析，同学们将经历真实数据分析的全流程、尝试解决真实数据分析中遇到的问题，从而在实践中加深对统计方法的理解与运用；还能以更全面、更批判性的视角看待前人研究并进行验证，为心理学的发展做出自己的贡献。

**2 进行可重复性检验**

**2.1 选取文献**

**2.1.1 与分析/计算可重复性相关要求**

为了实现检验原文献计算可重复性这一目的，在选取文献时应注意以下几点：

(1) 选定文献的数据和代码应公开提供并可免费获取，若无公开代码，至少可以获得可分析的数据；

(2) 对于选定文献所研究的领域，应有一定的背景知识储备，有相关的科研训练或实践更佳；

(3) 所选定文献中方法部分提到的数据分析或建模方法。

**2.1.2 与本课程相关的要求**

由于本课程的目标在于让同学们学习 R 编程语言及其在心理学研究中的运用，因此，文献选取为心理学各领域的量化研究的文献，在选取文献时还应注意：

(1) 选定文献应来自 *Science*、*Nature*、*Science Advances*、*Nature Human Behavior*、*Nature Communications*、*Psychological Science*、*Cognition* 、*Journal of Experimental Psychology: General*、*Advances in Methods and Practices in Psychological Science*、*Collabra: Psychology*这十本期刊中的人类心理与行为研究相关的文章(这些期刊均较为支持开放科学实践，更有可能提供了完整数据)；

(2) 选定文献的数据分析方法中至少有一部分是小组成员已经掌握或者通过快速学习能够掌握的方法。如果原文献使用R 语言完成分析并提供代码最佳；而使用R语言之外的工具但其方法能够在R语言中实现也可以纳入选择范围。

**2.2 进行分析/计算可重复检验**

在选定文献后，同学们由 4-5人为一组，以小组为单位对其结果进行分析/计算可重复性检验，即按照原文献报告的方法，对原文献的数据进行处理和分析，并将分析结果与原文献的结果进行比较。

以往对分析/计算可重复性的严谨研究中，通常会由多名重复者或者团队独立地检验同一篇文章以尽量避免操作错误(Crüwell et al., 2023)。首先，每个重复者会分别进行计算可重复性的检验并撰写个人报告，这一过程中重复者需要尽量保持独立、避免相互干扰；在这一阶段结束后，负责同一篇文章的重复者会就各自独立检验的结果进行交流、讨论，共同完成小组汇总报告。

但这种方式的工作难度和负担较大，本课程仅需要以小组为单位撰写报告，不要求每名成员撰写独立的个人报告，但如有兴趣，也推荐。在进行小组作业的过程中，推荐由各成员独立完成分析，后进行小组讨论的环节。在最终各小组汇报展示时，其他组同学也可以对汇报组的计算可重复检验的流程与结果提出意见建议。

**2.3 撰写计算可重复性检验报告**

**2.3.1 一般性内容**

如上文所述，同学们在完成对选取文献结果的计算可重复性检验后，需要撰写小组课程报告。为了帮助同学们更好地撰写报告，本文档根据已有可重复性检验研究的报告整理了可供本课程使用的报告格式(详见附件一)。同学们也可以在撰写报告的过程中，参考这些已发表的可重复性检验研究报告(Crüwell et al., 2023; OSF: https://osf.io/xzke7/)。但是，本文档中提供的模板并不是强制性的，大家也可进行重复分析之外的扩展分析。

课程报告应按照心理学研究中实证研究的规范进行撰写，包括标题、作者、摘要、关键词、引言、方法、结果、讨论与参考文献几个必要的部分。其中标题为《对*作者(年份)*研究结果的计算可重复性检验报告》，作者为小组成员及分工；前言部分报告所选文献以及对所选文献的简要介绍；方法部分介绍原研究的研究设计、数据分析以及数据来源（代码与数据的OSF 链接）等以及重复时所使用的数据、软件与分析思路；结果部分包括重复分析的结果及与原结果对比，讨论部分则对分析/计算可重复性的结果进行总结并探讨可重复或不可重复的原因。参考文献按照APA格式进行撰写。

**2.3.2 计算可重复性检验的思路与过程**

在介绍完选取的文献原来的方法后，方法部分可以撰写本小组重复检验时的思路和过程，包括以下几点：

(1) 指出检验了原文献哪部分的结果，如研究一、研究二/ 行为数据、电生理数据/ A任务、B任务等；

(2) 对使用的数据进行说明：原文献是否提供了研究中所用到的原始数据、处理后的数据和数据处理方法，检验时使用的为原始数据还是处理后的数据，若使用原始数据并对其进行处理，还需报告数据处理过程；

(3) 对使用的代码进行说明：原文献是否提供了研究中所用的代码以及关于代码的说明文件，检验时使用的为原代码还是自己编写的代码，若使用原代码则需记录原代码中出现的错误，若使用自己编写的代码则需要进行记录和说明。

**2.3.3 计算可重复性检验的结果\* (请注意阅读本小节)**

在每部分结果中，同时呈现重复分析结果与原文献报告的结果，并对两者比较。对于检验描述性统计[[1]](#footnote-1)的结果，需对样本量(*N*)、均值(Mean)、方差(*SD*)进行比较；对于检验推断性统计的结果，则需要对各类效应的显著性(*p*)、效应量，以及其他相关参数(如 *t* 值、*F*值、*r*值、*β*值、贝叶斯因子BF)进行比较。

从具体的步骤来看：

首先，逐一列出原文献报告的结果和和重复后的结果。

其次，计算可重复性系数*δ* (Hardwicke, 2021; Kambouris, 2024),*δ* = 100 × |XR − XO|/|XO|，其中 XR 为原文献报告的数值，XO 为重复分析时获得的数值。

最后，可根据计算的*δ* 对每个值的可重复性进行评级，等级包括：完全一致(*δ* = 0%)；偏差较小(0% < *δ* < 10%)；偏差较大(*δ* ＞ 10%)；因舍入导致的偏差(XR 和XO 之间的差距可能由保留小数位而导致，如XO 为1.50，XR 为1.51或1.52)；无法进行可重复检验(如没有提供所需的数据等)

对比较结果进行说明，可以汇总整理成表格，如下表所示：

**表 1 复现结果的描述性/推断性统计**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 研究一 | | | 研究二 | | |
| *N* | *p* | *η*2 | *N* | *p* | *η*2 |
| 原文献  报告结果 |  |  |  |  |  |  |
| 本研究 |  |  |  |  |  |  |
| *δ* |  |  |  |  |  |  |
| 评级 |  |  |  |  |  |  |

**2.4 分析/计算可重复性结果总结与讨论**

**2.4.1 计算可重复性结果总结**

在完成研究报告主体部分的基础上，对重复分析情况做总结分析。在此前，同学们对原文献各部分的结果进行了检验，得到了不同值的 *δ* 和评级。此时可以对值的*δ* 和评级进行汇总，报告达到各个评级的值的数量和在全体值中的占比，整理成表格，并在展示和文档中进行汇报。

表格如下所示：

**表 2 计算可重复性的评估表**

|  |  |  |
| --- | --- | --- |
| 可重复性情况 | 数量及占比 | |
| *N \** | *%* |
| 完全一致(*δ* = 0%) |  |  |
| 偏差较小(0% < *δ* < 10%) |  |  |
| 偏差较小(*δ* ＞ 10%) |  |  |
| 因舍入导致的偏差 |  |  |
| 无法进行可重复检验 |  |  |

\* 这里的*N*指的在重复分析中，对重复分析结果与原结果进行配对比较的次数。例如，原文仅进行了一个*t*检验，则*N*＝1；如果原文进行了一个2\*2的方差分析，并进行简单效应分析，则有可能有7个统计检验的数值：两个主效应，一个交互作用，四个可能的效应效应分析的*t*或者*F*值，因此*N*＝7。

若出现结果存在偏差或不能进行重复检验的情况，需要对此进行总结。首先，可以先说明文章中哪些结果由于什么样的原因无法进行重复性检验，因此不确定其结果是否能保持一致(如原文没有提供相关数据，导致重复的过程中断)；其次，可以指出描述性统计中有哪些结果出现了偏差(如样本人数、均值等)，由于这一部分数据可能直接用在后续的推断性统计中、影响后续结果的有效性，还应说明出现偏差的数据是否在后续分析中进一步使用了、用在了哪些分析中；最后，可以按照研究假设，依次对推断统计中有哪些结果出现了偏差进行报告，重新考虑原文献的研究假设是否可以证实。

**2.4.2 分析不一致结果的原因**

**表3 计算上（不）可重复的原因分析表**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **可能原因** | | **研究一** | **……** | **研究n** |
| **一般开放获取性问题** | 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； |  |  |  |
| 个别结果的微小差异，可能是由于印刷或复制粘贴错误； |  |  |  |
| 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； |  |  |  |
| 在文章的开放实践声明中对研究的模糊标记。 |  |  |  |
| **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)； |  |  |  |
| 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； |  |  |  |
| OSF上的数据存储问题，包括文件损坏或无法下载。 |  |  |  |
| **数据开放获取特定问题** | 没有提供原始数据； |  |  |  |
| 没有提供处理后的数据； |  |  |  |
| 没有提供数据处理过程的描述或代码。 |  |  |  |
| **代码开放获取特定问题** | 缺乏共享的分析代码或建模代码； |  |  |  |
| 软件包或软件版本的问题。 |  |  |  |
| **其他可能因素** | 文章出版年代； |  |  |  |
| 重复者此前是否有过 R 语言使用经验； |  |  |  |
| 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)。 |  |  |  |

在可重复性报告的讨论部分，需要对计算可重复性检验结果的影响因素进行探讨。小组成员可结合先前研究中常见的原因(Crüwell et al., 2023; Kambouris et al., 2024; Karathanasis et al., 2022)，对本小组重复结果进行讨论。表2中列举先前研究中的相关原因分析，大家可以以此为参考，对原文献进行分析，推测可能导致可重复性检验结果差异的原因。

**2.5 其他思考**

在讨论部分还可自由总结数据分析过程中除计算可重复性问题外的其他思考。如：

(1) 课程学习方面：如对于 R 语言学习或统计方法的思考；

(2) 科学研究方面：如对于研究方法的思考；

(3) 小组合作方面：如对于合作方式、协作能力的思考；

(4) 课程作业方面：课程作业的收获与建议。

**参考文献**

Crüwell, S., Apthorp, D., Baker, B. J., Colling, L., Elson, M., Geiger, S. J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D. S., Zaneva, M., & Brown, N. J. L. (2023). What’s in a badge? A computational reproducibility investigation of the open data badge policy in one issue of Psychological Science. *Psychological Science*, *34*(4), 513–522.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, *8*(341).

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., ... & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8), 180448.

Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, B. N., deMayo, B. E., Long, B., Yoon, E. J., & Frank, M. C. (2021). Analytic reproducibility in articles receiving open data badges at the journal Psychological Science: An observational study. *Royal Society Open Science*, *8*(1), 201494.

Kambouris, S., Wilkinson, D. P., Smith, E. T., & Fidler, F. (2024). Computationally reproducing results from meta-analyses in ecology and evolutionary biology using shared code and data. *PLOS ONE, 19*(3), e0300333. https://doi.org/10.1371/journal.pone.0300333

Karathanasis, N., Hwang, D., Heng, V., Abhimannyu, R., Slogoff-Sevilla, P., Buchel, G., Frisbie, V., Li, P., Kryoneriti, D., & Rigoutsos, I. (2022). Reproducibility efforts as a teaching tool: A pilot study. *PLOS Computational Biology, 18*(11), e1010615.

Kitzes, J., Turek, D., & Deniz, F. (2017). *The practice of reproducible research: Case studies and lessons from the dataintensive sciences*. Oakland: University of California Press.

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402.

附录一、研究报告格式

**对*作者(年份)*研究结果的计算可重复性检验**

小组成员分工

|  |  |  |  |
| --- | --- | --- | --- |
| 组长 |  | | |
| 组员 |  | | |
| 分工 | | | |
| 数据分析 |  | PPT 制作 |  |
| 文字报告制作 |  | PPT 展示 |  |

\* 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**：200 ～ 400字，包括背景、方法、结果与讨论/结论几个关键部分。

**关键词**：三到五个关键词，需要包括“计算可重复性”

1 引言

**1.1 所选文献**

文献：参考文献

数据及代码：OSF 链接

**1.2 文献介绍**

简要介绍所选文献的研究背景、主要研究问题及假设、研究结果和结论)

**2 方法**

**2.1 原研究方法简介**

介绍研究设计、数据分析的工具以及分析思路。

本小组所选取复现的哪些部分，并对选用的数据和代码做简单说明。

**2.2 复现思路与R包**

介绍对哪些分析进行复现，以及使用哪些R包。

**3 结果**

**3.1 描述性统计**

对原文献描述性统计进行重复的结果，并汇总表格：

**表 1 复现结果的描述性统计**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 研究一 | | | 研究二 | | |
|  | *N* | Mean | *SD* | *N* | Mean | *SD* |
| 原研究  报告结果 |  |  |  |  |  |  |
| 本研究 |  |  |  |  |  |  |
| *δ* |  |  |  |  |  |  |
| 评级 |  |  |  |  |  |  |

**3.2 推断性统计**

报告对原文献推断性统计进行重复的结果，并汇总表格：

**表 2 复现结果的推断性统计**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 研究一 | | | 研究二 | | |
|  | *N* | Mean | *SD* | *N* | Mean | *SD* |
| 原文献  报告结果 |  |  |  |  |  |  |
| 本研究 |  |  |  |  |  |  |
| *δ* |  |  |  |  |  |  |
| 评级 |  |  |  |  |  |  |

**3.3 对原文计算可重复性进行评估**

报告对原文献推断性统计进行重复的结果，并汇总表格：

报告原文献的值的评级分布情况，整理成表格，如下表所示：

**表 3 计算可重复性的评估表**

|  |  |  |
| --- | --- | --- |
| 可重复性情况 | 数量及占比 | |
| *N* | *%* |
| 完全一致(*δ* = 0%) |  |  |
| 偏差较小(0% < *δ* < 10%) |  |  |
| 偏差较小(*δ* ＞ 10%) |  |  |
| 因舍入导致的偏差 |  |  |

\* 这里的*N*指的在重复分析中，对重复分析结果与原结果进行配对比较的次数。例如，原文仅进行了一个*t*检验，则*N*＝1；如果原文进行了一个2\*2的方差分析，并进行简单效应分析，则有可能有7个统计检验的数值：两个主效应，一个交互作用，四个可能的效应效应分析的*t*或者*F*值，因此*N*＝7。

若出现不一致的情况，需要文字总结出现了哪些不一致。

**4 讨论**

**4.1 计算可重复性检验结果分析**

结合下表，对原文献进行分析，推测可能导致可重复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 4 计算上（不）可重复的原因分析表**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **可能原因** | | **研究一** | **……** | **研究n** |
| **一般开放获取性问题** | 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； |  |  |  |
| 个别结果的微小差异，可能是由于印刷或复制粘贴错误； |  |  |  |
| 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； |  |  |  |
| 在文章的开放实践声明中对研究的模糊标记。 |  |  |  |
| **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)； |  |  |  |
| 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； |  |  |  |
| OSF上的数据存储问题，包括文件损坏或无法下载。 |  |  |  |
| **数据开放获取特定问题** | 没有提供原始数据； |  |  |  |
| 没有提供处理后的数据； |  |  |  |
| 没有提供数据处理过程的描述或代码。 |  |  |  |
| **代码开放获取特定问题** | 缺乏共享的分析代码或建模代码； |  |  |  |
| 软件包或软件版本的问题。 |  |  |  |
| **其他可能因素** | 文章出版年代； |  |  |  |
| 重复者此前是否有过 R 语言使用经验； |  |  |  |
| 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)。 |  |  |  |

**4.2 其他思考**

可自由总结数据分析过程中除计算可重复性问题外的其他思考，也可以包括对本课的建议。

**参考文献(APA格式)**

1. 注意：应首先对这一部分所使用样本的描述性统计结果检验，再对其推断性统计结果进行检验，以确认推断性统计所使用的样本本身是否存在问题(如重复检验时发现，重复后的样本人数与原文献中报告的样本人数存在不一致，即使用这一样本能重复原研究的推断性统计的结果，结果的有效性仍然存疑)。 [↑](#footnote-ref-1)