arXiv:2305.04064v1 [stat.AP] 6 May 2023

多中心重复研究的贝叶斯样本量确定

Konstantinos Bourazas[1],电子邮箱：bourazas.konstantinos@ucy.ac.cy
Guido Consonni[2],电子邮箱：guido.consonni@unicatt.it Laura Deldossi[2],电子邮箱：laura.deldossi@unicatt.it[1]塞浦路斯大学数学与统计系及KIOS卓越研究与创新中心，塞浦路斯尼科西亚[2]意大利米兰圣心天主教大学统计科学系。

摘要

一场持续发酵的“可重复性危机”使生命科学到社会科学等众多学科领域的科学发现受到质疑。重复研究旨在检验已发表研究中发现的有效性，并尝试评估这些发现与重复研究结果在统计上是否一致。尽管大多数重复研究项目基于单次实验，但在不同研究中心同时开展同一实验的多项独立重复正日益普遍。针对这类项目，我们研究如何检验研究中心之间的异质性；具体而言，我们着重探讨如何确定样本量，以便在收集实验数据后获得有力证据。

关键词：分析先验、贝叶斯设计、贝叶斯因子、设计先验、异质性。

1 引言

过去几十年来，人们逐渐发现，相当一部分已发表的结果无法重复验证（Ioannidis，2005），由此导致实证科学陷入“可重复性危机”（Baker，2016）。这一

现象已在社会科学领域（尤其是心理学和经济学；参见 Hensel，2021 等）以及生命科学领域（Zwanenburg，2019）得到指出和研究。造成这一现象的原因有很多，从发表偏倚（Francis，2012），到实验设计不佳（Pashler 和 Harris，2012），再到存疑的统计方法（Wasserstein 和 Lazar，2016）。

Simons（2014）指出，“可重复性是科学的基石”，因为它确立了已发表研究的有效性。多年来，已有许多关于重复研究分析的贡献；例如 Bonett（2012）、Zwaan 等人（2018）、Hedges 和 Schauer（2019a）、Hou 等人（2020）。贝叶斯视角下的研究包括 Wagenmakers 等人（2015）、Etz 和 Vandekerckhove（2016）以及 Marsman 等人（2017）。同样在贝叶斯方法中，Verhagen 和 Wagenmakers（2014）提出了重复贝叶斯因子，Harms（2019）将其应用于固定效应方差分析设计。此外，Pawel 和 Held（2020）从预测视角研究了重复研究，而 Held（2020）和 Held 等人（2022）则利用逆贝叶斯思想评估重复成功。最后，Muradchanian 等人（2021）对用于评估重复成功的频率学派和贝叶斯指标进行了比较研究。

大多数重复研究（RS）指的是单次重复研究，但人们对同时开展多项重复研究的兴趣日益增长，与此同时，Mathur 和 VanderWeele（2020）提出了新的指标。在这种情形下，分析往往聚焦于研究间变异，尤其是在多个实验重复之间的异质性问题备受关注时；参见 Klein（2014）以及 Hedges 和 Schauer（2019a、2019b 和 2021）。在更一般的框架下，Gronau 等人（2021）提出了一种贝叶斯模型平均程序，用于检验效应量是否存在，以及异质性是否存在。

值得强调的一个重要方面是，研究很少

会就设计选择提供统计学论据，尤其是设计是否足以在事前得出令人信服的结论，例如就检验功效（假设检验）或标准误（估计）而言。样本量确定（SSD）涉及研究中纳入的位点数量以及每个位点内的受试者数量，是成功开展重复研究设计的一项重要议题。Wong 等人（2021）和 Bonett（2021）回顾了重复研究的设计与分析，而 Simon（1999）以及 Bayarri 和 Mayoral（2002a 和 2002b）则研究了如何采用贝叶斯方法设计重复研究。一般而言，重复研究的设计可分为两类：受约束的设计，即将原始研究纳入考量，通常将问题表述为比较原始研究与后续重复研究；或不受约束的设计，即不将原始研究纳入分析。Hedges 和 Schauer（2021）提出了一种用于样本量确定的不受约束的多地点设计；另可参阅 Fedorov 和 Jones（2005）以及 Harden 和 Friede（2018）关于临床试验的研究。

本文探讨一种用于多地点重复实验的贝叶斯无约束设计，以检验研究中心之间是否存在异质性，并提出一种使用贝叶斯因子（BF）作为证据度量的 SSD 方法（Kass 和 Raftery，1995年）。可以将贝叶斯因子视为频率学派假设检验中似然比检验的贝叶斯对应方法，其中使用的是边际似然，而非最大化似然。贝叶斯因子的一个重要特点是，它可以比较任意一对假设（模型），因此不受限于嵌套模型。此外，它能够分别为两个相互竞争的假设提供证据；而在频率学派方法中，不显著结果无法被解释为支持零假设。有关贝叶斯因子及其用途的进一步讨论，请参见 Dienes（2014）和 Hoijtink 等人（2019）。

本文结构如下。在第2节中，我们提出一个能够解释异质性的层级模型，并识别代表无变异的子模型

；随后，我们讨论贝叶斯 SSD 中分析 先验与设计 先验之间的重要区别，推导用于检验研究中心间异质性的贝叶斯因子，并展示如何利用适当的算法近似其先验预测分布。在第3节中，我们根据贝叶斯因子及相应的先验概率介绍三类证据，随后提出 SSD 准则，并在若干为便于与替代方法比较而选定的情景下，得出最优样本量。最后，第4节进行简短讨论，重点指出一些值得进一步研究的问题。贝叶斯因子推导的技术细节见附录。关于替代模拟情景的敏感性分析和进一步结果，可在本文的在线补充材料中获取，另附 R-代码。

2 模型、先验分布与贝叶斯因子

本节介绍一种适用于描述异质性的层次模型，以及用于推导检验各研究中心间变异性的贝叶斯因子的分析方法。此外，我们讨论了在竞争假设下获得贝叶斯因子（先验）预测分布所需的设计先验，最后介绍两种从其预测分布中模拟贝叶斯因子取值的算法。

2.1 用于考虑异质性的层次模型

考虑 m 个相互独立的研究中心，并令 t_j 表示研究中心 j 的效应量估计量 $j = 1$，2，.....，m。采用荟萃分析框架，我们假设 t_j近似服从正态分布，其中心为研究中心特定的效应量μ_j，方差为 $\sigma_j^2 = \sigma^2/n_j$，其中 σ^2 为单位方差。当样本量较大时，可以认为 σ^2 已知，因为它

可以被准确估计。为简化 SSD 的阐述，我们假设采用完全平衡设计（例如参见 Fedorov 和 Jones，2005），即每个研究中心招募相同数量的受试者，因此 $n_j = n, \quad \forall j$。对于每个 $j = 1, \ldots\ldots, m$，我们分别考虑层次模型。

$$\begin{aligned} t_j|\mu_j &\sim N(\mu_j, \sigma^2/n) \\ \mu_j|\mu, \tau^2 &\sim N(\mu, \tau^2), \end{aligned} \tag{1}$$

其中，μ 是总体平均效应量，τ^2 表示研究中心之间的异质性；有关基于有向无环图（DAG）的可视化示意，请参见图1。我们将在第2.2小节讨论 μ 和 τ^2 的先验分布。

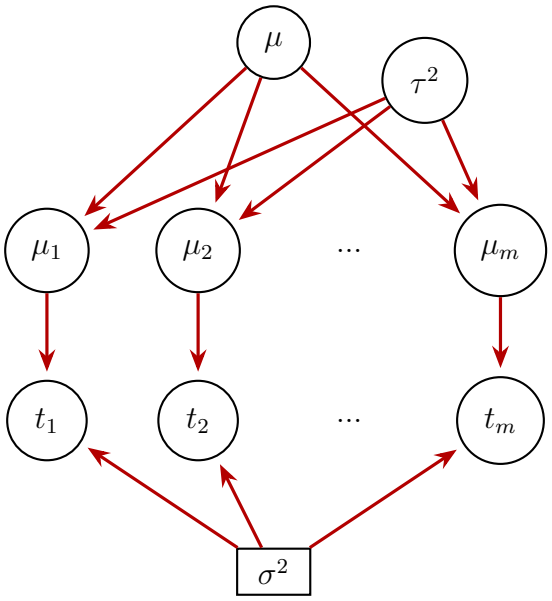


图1：层次模型（1）的DAG

对（1）中的 μ_j 进行积分，得到

$$\mathcal{M}_1 : t_j|\mu, \tau^2 \sim N(\mu, \tau^2 + \sigma^2/n). \tag{2}$$

将 $\tau^2 = 0$ 代入（2），得到无异质性模型

$$\mathcal{M}_0 : t_j|\mu \sim N(\mu, \sigma^2/n). \tag{3}$$

我们将在第2.3小节中使用贝叶斯因子（BF）比较模型 $\mathcal{M}_0$ 和 $\mathcal{M}_1$ 。

2.2 先验分布

首先考虑 τ^2。结合单位方差 σ^2，可以更好地理解其可能的取值范围。因此，换回标准差后，我们以相对异质性 $\gamma = \tau/\sigma$为单位进行分析。

请记住，实验设计是一项前瞻性事业，旨在数据到来之前达到预期水平的推断性能。这意味着必须确定一个足够的样本量。不同于以待检验参数 γ 的固定值为条件的频率学派功效分析，贝叶斯方法要求在参数空间上指定完整的先验分布。在此背景下，通常区分两类先验分布：分析先验，记为 $h_a(\gamma)$；以及设计先验，记为 $h_d(\gamma)$；参见 O'Hagan 和 Stevens（2001）以及 O'Hagan 等（2005），他们在临床试验背景下使用了这两类先验。分析先验用于数据到来后进行推断；在我们的案例中，它将用于评估贝叶斯因子。原则上，它应当是弱信息先验，从而能够被广泛接受，同时又应当是适当先验，以便能够明确地评估贝叶斯因子。另一方面，设计先验应当是信息先验，代表研究者对于预期或认为值得检测的异质性大小的看法。

我们首先考虑分析先验 $h_a(\gamma)$。根据 R¨over 等人（2021）对异质性弱信息先验所做的广泛综述，我们假设其服从 Half-$t(\nu_\gamma,\ \sigma_\gamma)$ 分布，其中 ν_γ 是自由度

，σ_γ 是尺度参数。Half-t(ν_γ, σ_γ) 分布是以零为中心、自由度为 ν_γ 、尺度为 σ_γ 的 Student-t 变量绝对值的分布。当 ν_γ取较小值时，其密度单调递减，且上尾较厚。对于超参数的选择，我们设定 $\nu_\gamma = 4$；参见 R¨over 等人（2021）。此外，我们设定 $\sigma_\gamma = 1/7$，使得 Half-t 分布的 95 % 分位数约为 0.4，这一数值大体符合 Hedges 和 Schauer（2021）针对多个应用领域提出的建议。

现在我们来讨论设计先验 $h_d(\gamma)$，并将其设为折叠 t 分布（Psarakis 和 Panaretoes，1990）。折叠 t 分布是一个变量的绝对值的分布，该变量服从位置参数为 μ_γ、尺度为 σ_γ 、自由度为 ν_γ 的非标准化 t 分布；当 $\mu_\gamma = 0$ 时，它便简化为半 t 分布。关于超参数，我们令 $\nu_\gamma = 4$ 并设定 $\mu_\gamma = 0.2$；后一种选择基于对相对异质性 γ 的合理取值所做的评估，相关讨论见 Hedges 和 Schauer（2021）。在位置参数周围加入适度的不确定性，例如 ± 0.05，并调整 $\sigma_\gamma = 1/55$ ，使区域 $[0.15, 0.25]$ 的可信区间覆盖率达到 95%。图 2 展示了当前设定中分析所用先验和设计先验的密度图。补充材料对其他超参数配置进行了敏感性分析。

现在考虑同时出现在模型 $\mathcal{M}_1$ 和 $\mathcal{M}_0$ 中的 μ 的先验；分别参见公式（2）和（3）。在检验异质性时，总体均值是一个干扰参数，我们假设它在先验上与 γ 独立。我们建议使用杰弗里斯先验 $\pi(\mu) \propto 1$。尽管它是不适当的先验，但在我们的情形下，它适用于推导贝叶斯因子，因为 μ 是两个模型共有的参数。

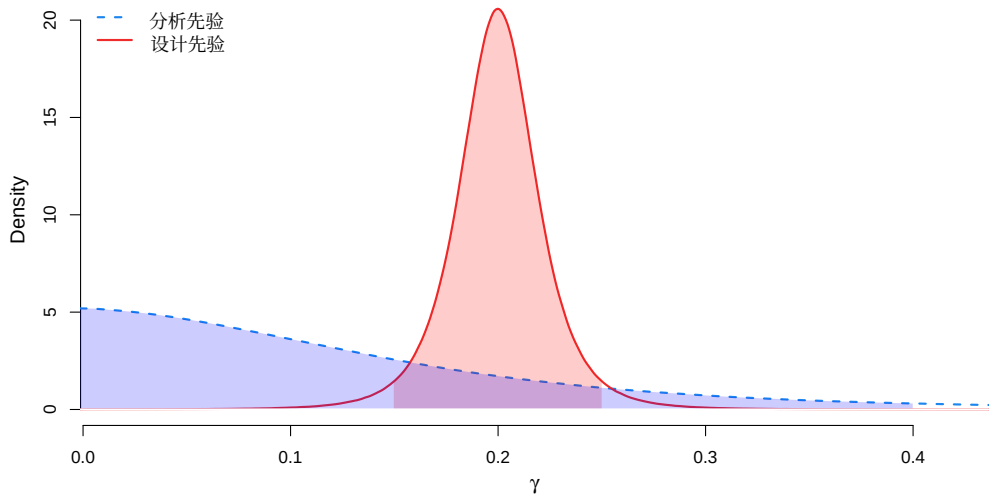


图 2：分析先验 $h_a(\gamma)$（虚线）和设计先验 $h_d(\gamma)$（实线），以及它们的 95% 可信区间（高亮区域）。

2.3 贝叶斯因子及其预测分布

我们通过贝叶斯因子 BF_{01} 比较模型 $\mathcal{M}_0$ 和 $\mathcal{M}_1$，该因子是模型 $\mathcal{M}_0$ 和 $\mathcal{M}_1$ 下边际数据分布的比值。度量 BF_{01} 量化了零假设相对于备择模型所获得的支持；例如，数值 $BF_{01} = 3$ 表示 $\mathcal{M}_0$ 比 $\mathcal{M}_1$ 的可能性高三倍。Jeffreys（1961）提出了一种启发式分类方案，将数值归入若干类别，以解释贝叶斯因子所提供的证据；后续的阐述可参见 Kass 和 Raftery（1995）以及 Sch¨onbrodt 和 Wagenmakers（2018）。

根据 (3) 和 (2) 可得下述边际数据分布

模型 $\mathcal{M}_0$，分别为 $\mathcal{M}_1$，

$$m_0(\boldsymbol{t_r}) = \int_{-\infty}^{+\infty} f(\boldsymbol{t_r}|\mu, \mathcal{M}_0)\pi(\mu)d\mu = \left(\frac{1}{m}\right)^{1/2} \cdot \left(\frac{2\pi\sigma^2}{n}\right)^{(1-m)/2} \cdot \exp\left\{-\frac{Q}{2}\right\}, \tag{4}$$

$$\begin{aligned} m_1(\boldsymbol{t_r}) &= \int_0^{+\infty}\int_{-\infty}^{+\infty} f(\boldsymbol{t_r}|\mu, \gamma, \mathcal{M}_1)h_a(\gamma)\pi(\mu)d\mu d\gamma \\ &= \left(\frac{1}{m}\right)^{1/2} \cdot \left(2\pi\sigma^2\right)^{(1-m)/2} \int_0^{+\infty} \left(\frac{1}{n} + \gamma^2\right)^{(1-m)/2} \exp\left\{-\frac{Q}{2} \cdot \left(1 + n\gamma^2\right)^{-1}\right\} h_a(\gamma)d\gamma, \end{aligned} \tag{5}$$

其中 $\boldsymbol{t_r} = (t_1, \ldots\ldots, t_m)$，$\bar{t}$ 是 t_j 的样本均值，$Q = n \cdot \sum_{j=1}^{m}(t_j - \bar{t})^2 / \sigma^2$ 是一个其分布不依赖于 σ^2 的量。可以看出，$m_0(\boldsymbol{t_r})$ 可以用闭式表示。另一方面，$m_1(\boldsymbol{t_r})$ 的计算可以使用蒙特卡罗模拟进行近似。所得贝叶斯因子为:

$$BF_{01}(\boldsymbol{t_r}) = \frac{m_0(\boldsymbol{t_r})}{m_1(\boldsymbol{t_r})} = \frac{n^{(m-1)/2} \cdot \exp\left\{-\frac{Q}{2}\right\}}{\int_0^{+\infty} \left(\frac{1}{n} + \gamma^2\right)^{(1-m)/2} \cdot \exp\left\{-\frac{Q}{2} \cdot (1 + n\gamma^2)^{-1}\right\} h_a(\gamma)d\gamma}, \tag{6}$$

它仅通过 Q依赖于数据。我们强调，在设计阶段，观测值 $\boldsymbol{t}_r$ 尚不可用。因此，若要基于贝叶斯因子获得明确无歧义的结果，就需要 $BF_{01}(\boldsymbol{t_r})$的先验预测分布，而这又取决于 Q的（先验）预测分布。以下是两个众所周知的事实：i) 在 $\mathcal{M}_0$条件下，$Q \sim \chi^2_{m-1}$，其中 χ^2_p 表示自由度为 p 的卡方分布；ii)

在 $\mathcal{M}_1$条件下，给定 γ时，$(1+n\gamma^2)^{-1}\cdot Q \sim \chi^2_{m-1}$ （Hedges 和 Pigott, 2001）。要获得 Q 在 $\mathcal{M}_1$条件下的无条件分布，还需要对 $\gamma \sim h_d(\gamma)$进行进一步混合。最后，要从其先验预测分布中获得$BF_{01}(\boldsymbol{t_r})$的一个实现值，必须将上述事实 i) 和 ii) 与对式 (6) 分母中积分的求值结合起来。上述计算过程将采用蒙特卡洛近似完成。算法 1 描述了假设模型为 $\mathcal{M}_0$时的步骤，而算法 2 则适用于 $\mathcal{M}_1$。

算法 1： 从.....的先验预测分布中模拟
BF_{01} 在 $\mathcal{M}$条件下 0

输入： 研究中心数 m，每个位点的受试者数 n，随机
从分析先验 $\boldsymbol{\gamma}_a$生成的大小为 S 的样本，
迭代次数 T

输出： 来自.....预测分布的大小为 T 的向量 $\boldsymbol{BF}_{01}$
贝叶斯因子$_{01}$ 在 $\mathcal{M}$下 0

对于 $t = 1, \cdots, T$ **do**
  生成 $q \sim \chi^2_{m-1}$
  $g_0 \leftarrow n^{(m-1)/2} \cdot$ 指数 $\{-q/2\}$
  $g_1 \leftarrow \dfrac{1}{S}\sum_{j=1}^{S}\left[1/n+\gamma^2_{a,j}\right]^{(1-m)/2}\cdot \exp\left\{-q/2\cdot\left(1+n\gamma^2_{a,j}\right)^{-1}\right\}$
  $BF_{01,t} \leftarrow g_0/g_1$
结束

图3报告了在模型 $\mathcal{M}_0$ （上方面板）和 $\mathcal{M}_1$ （下方面板）下，若干组乘积 $n\cdot m$ 保持不变的 (n, m) 的 BF 先验预测分布（以对数刻度表示）。具体而言，在左侧两个面板中，我们设定 $n = 80$，其中 $m \in \{4,\ 8,\ 12\}$；而在右侧面板中，我们固定 $m = 8$，其中 $n \in \{40,\ 80,\ 120\}$。按行阅读（即针对固定的真实模型），左侧的每个分布都对应右侧一个具有相同 $n\cdot m$ 的分布。这样便能更好地看出，如何在研究中心数量 m 和受试者总数之间重新分配给定的受试者总数所产生的影响。

算法 2： 从以下先验预测分布中模拟 BF_{01} 在 $\mathcal{M}$下 $_1$

输入： 研究中心数量 m，每个位点的受试者数量 n，随机从分析先验 $\boldsymbol{\gamma}_a$生成的大小为 S 的样本，迭代次数 T，从设计先验 T 生成的大小为 T 的随机样本从设计先验 $\boldsymbol{\gamma}_d$中

输出： 来自.....预测分布的大小为 T 的向量 $\boldsymbol{BF}_{01}$

贝叶斯因子$_{01}$ 在 $\mathcal{M}$下 $_1$

对于 $t = 1, \ldots, T$ **do**

  生成 $q \sim \chi^2_{m-1}$

  $q' \leftarrow q' \cdot (1 + n\gamma^2_{d,t})$

  $g'_0 \leftarrow n^{(m-1)/2} \cdot \exp\{-q'/2\}_{(1}{}_{m)/2} \cdot \exp\left\{-q'/2 \cdot \left(1 + n\gamma^2_{a,j}\right)^{-1}\right\}$

  $g'_1 \leftarrow \frac{1}{S}\sum_{j=1}^{S}\left[1/n + \gamma^2_{a,j}\right]$

  $BF_{01,\ t} \leftarrow g'_0/g'_1$

结束

每个位点的受试者人数 n。为便于可视化，突出显示了配对（n，m）$=$（80，8）的分布（ $\mathcal{M}_0$ 下为蓝色， $\mathcal{M}_1$下为红色），因为它们在两种设置中都很常见。

图3揭示了在两个模型下贝叶斯因子分布学习速率之间的显著不平衡。显然，在 $\mathcal{M}_0$ 下，最大log(BF_{01})（代表支持真实模型的证据）从未达到10；而在 $\mathcal{M}_1$ 下，相应的证据可达到 10^6 甚至更高（回想一下，$BF_{10} = 1/BF_{01}$，其中 BF_{10} 衡量支持 $\mathcal{M}_1$的证据）。

Dawid（2011）从理论角度研究了这一现象，基本证明了对于嵌套模型（如我们的设定），在零假设模型下，贝叶斯因子随样本量的平方根增长，而在较大的包容模型下，其增长则呈指数级；另见 Johnson 和 Rossell（2010）。然而，这一结果仍然

在论文中，这一问题相对较少受到关注，少数值得注意的例外包括 Sch¨onbrodt 和 Wagenmakers（2018年）以及 Ly 和 Wagenmakers（2022年）。

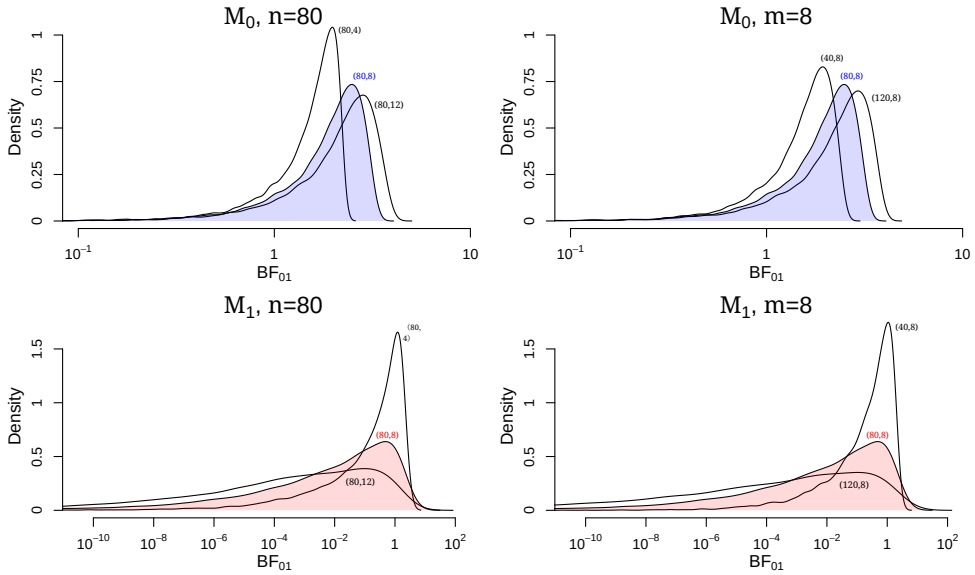


图3：在 $\mathcal{M}_0$ （上排）和 $\mathcal{M}_1$（下排）条件下，所选 $(n,\ m)$ 组合的贝叶斯因子先验预测分布，其中 $n \in \{40,\ 80,\ 120\}$ 和 $m \in \{4,\ 8,\ 12\}$。

3 样本量确定

本节考察真实模型为 $\mathcal{M}_0$ 或 $\mathcal{M}_1$ 时，正确证据、误导性证据和未确定证据的概率。基于这些概率，我们提供一个设计框架，用于确定配置，以便针对异质性进行检验时，配对 $(n,\ m)$ 能够提供有力证据。

3.1 贝叶斯因子阈值与模型证据分类

对于给定的正阈值 k_0 和 k_1，如果 BF_{01} 大于 k_0，则数据表明支持 $\mathcal{M}_0$ 的证据（水平为 k_0）；而如果它小于 $1/k_1$，则数据表明支持 $\mathcal{M}_1$ 的证据（水平为 k_1）。最后，如果贝叶斯因子位于区间 $(1/k_1,\ k_0)$ 内，则证据未确定。如果将 k_0 设为 10 或更高等较大值，那么 $BF_{01} > k_0$ 可被视为支持 $\mathcal{M}_0$ 的强 证据（Kass 和 Raftery，1995年；Sch¨onbrodt 和 Wagenmakers，2018年），或者用 De Santis（2004年）的话说，是支持 $\mathcal{M}_0$ 的决定性 证据。对于 $\mathcal{M}_1$ 的证据，k_1 也适用类似的考虑。相反，较小的 k_0 和 k_1 值，例如 3 或 5（有时会采用这些值；Weiss，1997年），只能表明任一方向上的中等 证据。请注意，允许使用不同的阈值 k_0 和 k_1，通常 $k_0 < k_1$，因为在 $\mathcal{M}_0$ 下学习真实模型的速度比在 $\mathcal{M}_1$ 下更慢；请参见图3和图4，以及我们在第2.3小节末尾的评论。

与其预先将 k_0 和 k_1 固定为证据阈值，不如根据设计方面的考量（例如第一类错误的概率 α 和检验功效（$1-\beta$），二者均从贝叶斯视角进行解释）间接地设定它们。这样一来，贝叶斯因子仅仅充当检验统计量，其固有含义以及对证据界限 k_0 和 k_1的实质性解释也就不复存在。另一方面，由此得到的界限值能够满足更为传统的设计目标，因此可能更容易被实践者广泛接受。

给定阈值 k_0 和 k_1，并假设 $\mathcal{M}_0$ 或 $\mathcal{M}_1$ 依次为真实的数据生成模型，我们定义以下事件（为简洁起见，省略其对所选阈值的依赖关系）：

- 正确 证据（C）：证据支持正确的模型
- 误导性证据（M）：证据支持错误的模型

- 未定（U）证据： C 和 M 均不成立。

在假设上述任一模型成立的情况下，自然可以评估上述各事件的条件（先验预测）概率。分别在 $\mathcal{M}_0$ 和 $\mathcal{M}_1$的假设下应用算法 1 和 2，即可对这些概率进行近似计算。还可以使用先验模型概率 $\pi_0 = p(\mathcal{M}_0)$ 和 $\pi_1 = p(\mathcal{M}_1) = 1 - \pi_0$，对两个模型下相应的条件概率取平均，从而评估这些事件 $\{C, M, U\}$ 的无条件或总体概率；汇总见表 1。

图 4 显示了 BF_{01} 在 $\mathcal{M}_0$ （左图）和$\mathcal{M}_1$ （右图）下的分布，并突出显示了在（n, m） = （80,8）和 $k_0 = k_1$ =3 条件下正确、误导和未定证据的相应概率。显然，在 $\mathcal{M}_1$ （78%）下，正确证据的条件概率明显高于在 $\mathcal{M}_0$ （4%）下的条件概率；相应地，在 $\mathcal{M}_0$ （94%）下，未定证据的条件概率高于在 $\mathcal{M}_1$（21%）下的条件概率。这进一步表明，在 $\mathcal{M}_0$ 下学习确实更加困难。

表 1：正确、误导性和未定证据的概率。

	正确证据	误导性证据	未定证据
$\mathcal{M}_0$	$p_0^C = p\ (BF_{01} > k_0 \mid \mathcal{M}_0)$	$p_0^M = p\ (BF_{01} < 1/k_1 \mid \mathcal{M}_0)$	$p_0^U = p\ (1/k_1 < BF_{01} < k_0 \mid \mathcal{M}_0)$
$\mathcal{M}_1$	$p_1^C = p\ (BF_{01} < 1/k_1 \mid \mathcal{M}_1)$	$p_1^M = p\ (BF_{01} > k_0 \mid \mathcal{M}_1)$	$p_1^U = p\ (1/k_1 < BF_{01} < k_0 \mid \mathcal{M}_1)$
总体	$p^C = \pi_0 \cdot p_0^C + \pi_1 \cdot p_1^C$	$p^M = \pi_0 \cdot p_0^M + \pi_1 \cdot p_1^M$	$p^U = \pi_0 \cdot p_0^U + \pi_1 \cdot p_1^U$

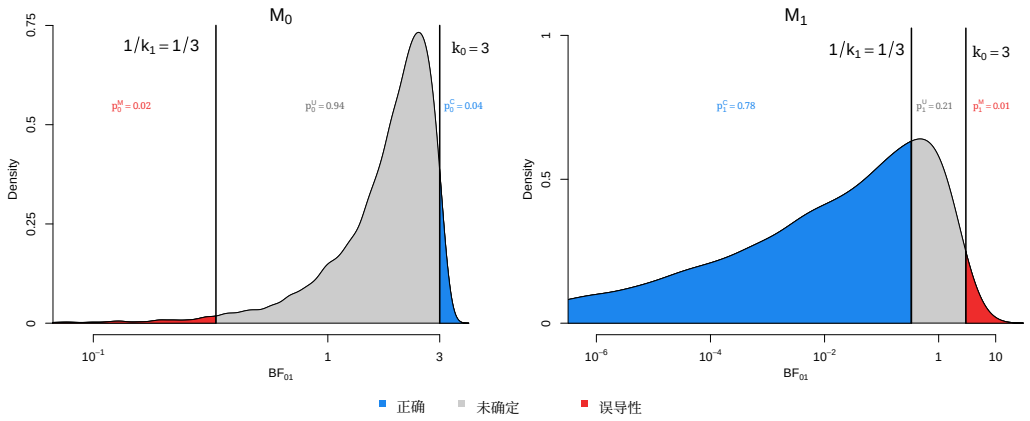


图 4：在 $\mathcal{M}_0$和 $\mathcal{M}_1$下，$(n,\ m) = (80,\ 8)$ 中正确、误导性和未确定证据的概率。

3.2 样本量确定的条件方法

3.2.1 样本量确定的条件判据

在条件SSD 方法中，通常要求在感兴趣参数的给定取值下，确保达到期望的检验功效水平（通常为 0.8 或更高）。在贝叶斯框架下，固定取值被整个分布取代，即设计先验，从而得到无条件检验功效（但仍取决于先验本身）。Weiss（1997）使用贝叶斯因子作为证据的度量，探讨了基于第一类错误率以及条件和无条件检验功效的假设检验样本量确定（当时尚未区分分析先验和设计先验）。Wang 和 Gelfand（2002）提出了其他准则，用于选择样本量，以便在具备合理先验保证的情况下区分两个模型。

在本文关注的多地点重复研究设定中，检测异质性是否存在（模型 $\mathcal{M}_1$）至关重要。因此，我们特别关注 p_1^C，即正确识别 $\mathcal{M}_1$的概率，因为它最为相关。这引出了以下设计准则

针对给定数量的研究中心 m，最优选择受试者数量 n^* 的准则：

$$n^* = \min\{n \in \mathbb{N} : p_1^C \geq (1-\beta) \ \ and \ \ p_0^M = \alpha\}, \tag{7}$$

其中， $\mathbb{N}$ 是自然数集合，$(1-\beta)$ 是成功检测出异质性的无条件检验效能，而 α 是第一类错误率。

实际上，对于固定值 m ，要得到 n^* ，可以按如下步骤进行。固定 m（为简洁起见，记号中省略），令 $n = n_0$ ，并令 $1/k_1^{\alpha}(n_0)$ 等于在 $\mathcal{M}_0$下 $BF_{01}(n_0)$ 分布的 α分位数，使得 $p_0^M(n_0) = p(BF_{01}(n_0) < 1/k_1^{\alpha}(n_0)|\mathcal{M}_0) = \alpha$。接着计算 $p(BF_{01}(n_0) < 1/k_1^{\alpha}(n_0)|\mathcal{M}_1)$；如果该值小于$(1-\beta)$，则将 n 增至 $n_1 > n_0$。从图3可以看出，在 $\mathcal{M}_0$ 下，$BF_{01}(n)$ 的分布会随着 n 增大而向右移动（ m固定时）。非正式地说，为满足约束 $p_0^M(n_1) = p(BF_{01}(n_1) < 1/k_1^{\alpha}(n_1)|\mathcal{M}_0) = \alpha$，我们可以据此得出结论： $k_1^{\alpha}(n_1) < k_1^{\alpha}(n_0)$。因此， $p(BF_{01}(n_1) < 1/k_1^{\alpha}(n_1)|\mathcal{M}_1) > p(BF_{01}(n_0) < 1/k_1^{\alpha}(n_0)|\mathcal{M}_1)$，因为在 $\mathcal{M}_1$ 下， $BF_{01}(n)$ 的分布会随着 n 增大而向左移动。如果 $p(BF_{01}(n_0) < 1/k_1^{\alpha}(n_0)|\mathcal{M}_1) > (1-\beta)$，则直接选择 $n_1 < n_0$。随后可以通过假位法（False Position method）（Burden 等人，2015），并作适当修改以考虑 n的离散性，求得最优解 n^*。用于计算式（7）中 n^* 的 R-代码可在补充材料中找到。

3.2.2 条件方法的结果

在本小节中，我们在四种情景下实施条件方法，这四种情景对应于 $p_1^C \in \{0.80, 0.90\}$ 和 $p_0^M \in \{0.01, 0.05\}$的不同组合；每种情景均针对研究中心数量的十五个可能取值进行分析$m \in \{3, 4, ..., 17\}$。由于我们关注的是研究中心之间的变异性，因此我们排

表2： $p_1^C \in \{0.8, 0.9\}$ 和 $p_0^M \in \{0.01, 0.05\}$的数对集合（n^*， m），以及条件方法中对应的阈值 $1/k_1$。

$p_1^C = 0.8$				$p_1^C = 0.9$			
$p_0^M = 0.01$		$p_0^M = 0.05$		$p_0^M = 0.01$		$p_0^M = 0.05$	
(n^*, m)	$1/k_1$	(n^*, m)	$1/k_1$	(n^*, m)	$1/k_1$	(n^*, m)	$1/k_1$
(518, 3)	0.220	(328, 3)	0.639	(1138, 3)	0.253	(730, 3)	0.756
(268, 4)	0.214	(178, 4)	0.603	(478, 4)	0.234	(323, 4)	0.675
(184, 5)	0.215	(126, 5)	0.591	(302, 5)	0.230	(211, 5)	0.649
(143, 6)	0.215	(99, 6)	0.584	(221, 6)	0.227	(157, 6)	0.634
(117, 7)	0.213	(82, 7)	0.577	(176, 7)	0.223	(127, 7)	0.623
(100, 8)	0.215	(71, 8)	0.575	(147, 8)	0.224	(107, 8)	0.616
(88, 9)	0.212	(62, 9)	0.582	(127, 9)	0.220	(92, 9)	0.622
(79, 10)	0.214	(56, 10)	0.582	(113, 10)	0.222	(82, 10)	0.619
(72, 11)	0.213	(51, 11)	0.580	(101, 11)	0.219	(74, 11)	0.617
(67, 12)	0.210	(48, 12)	0.582	(93, 12)	0.216	(68, 12)	0.613
(62, 13)	0.213	(45, 13)	0.581	(85, 13)	0.219	(63, 13)	0.610
(58, 14)	0.215	(42, 14)	0.579	(79, 14)	0.220	(59, 14)	0.613
(55, 15)	0.215	(40, 15)	0.581	(74, 15)	0.220	(55, 15)	0.609
(52, 16)	0.217	(38, 16)	0.580	(69, 16)	0.221	(52, 16)	0.609
(49, 17)	0.217	(36, 17)	0.581	(66, 17)	0.222	(49, 17)	0.607

除了取值 $m = 2$，因为该取值也会导致 n大幅升高；另请参见图5。

μ 和 γ 的先验分布选择方式如第2.2节所述；此外，在算法 1 和 2 中，我们设定 $S = 10{,}000$ 和 $T = 50{,}000$。结果见表2，其中阈值 $1/k_1$ 按配对（n^*, m）报告。相同的结果也可在图5中直观查看。显然，对于每个 m，对异质性正确识别概率 p_1^C提出更高的要求，会产生更高的值 n^*；降低 p_0^M 时也会出现相同情况。对于给定的 p_0^M，阈值 $1/k_1$ 似乎对p_1^C的选择以及不同配置（n^*, m）均具有稳健性，这反映出在所考察的取值范围内， BF_{01} 分布的下尾敏感度较低。请注意，将p_0^M从 0.05 降低至 0.01（减少五倍），会使 $1/k_1$ 大约减少三倍（约从 0.6 降至 0.2）。在证据尺度上

支持 $\mathcal{M}_1$，即 $BF_{01} < 1/k_1$。根据 Sch¨onbrodt 和 Wagenmakers（2018年）的研究，这意味着证据等级从“轶事性”（$1/3 < BF_{01} < 1$）提升至“中等”（$1/10 < BF_{01} < 1/3$）证据。这表明，较高 n^* 所固有的更高成本可能值得付出，不仅是为了实现更低的第一类错误率，也是为了在异质性实际存在时获得更有说服力的证据来检测异质性。另一个值得提及的特点是 m 与 n之间的相互作用。鉴于二者呈反向关系，可以发现，只有在 m 取值较小时，n 才会出现特别显著的下降，此后下降幅度则相对温和；参见图5。最优选择 n^* 也对设计 先验敏感。如果后者趋近于零（即更接近零假设模型），则需要更高的 n^*。具体而言，在保持 ν_γ 和σ_γ 不变，并考虑到 $\mu_\gamma = 0.2$的情况下，如果我们转向 $\mu_\gamma = 0.1$，则平均所需样本量将增加至原来的四倍；而如果设定为 $\mu_\gamma = 0.3$，样本量则可以减半。模拟结果已通过解析方式列于补充材料A中。

将我们的结果与 Hedges 和 Schauer（2021，表2）中的结果进行比较颇有启发。请注意，他们的论文引入了一个成本函数，因此其结果取决于 c_2/c_1，即每个实验室的成本与每个受试者的成本之比。他们针对 c_2/c_1 取五个值构成的网格，并对每个值，针对选定的相对方差异质性数值（τ^2/ω 是他们的记号），确定最优样本量（n_O，m_O）；该异质性对应于我们的 γ^2。以上分析在两个（条件）检验功效水平下重复进行，即 0.8 和 0.9。为将他们的结果与我们的结果进行比较，我们首先确定相对方差的一个值 $\gamma_0^2 = 0.04$，使其大致“匹配”设计先验中的先验期望 $E(\gamma) = 0.2$。接下来，我们针对每个 $c_2/c_1 = r_k$，在 $k = 1$，.....，5 的条件下，计算最优样本量的集合（n_{Ok}，m_{Ok}）。对于给定的 m_{Ok}，将我们的最优 n_k^* 与 n_{Ok} 进行比较可见，前者至多仅比后者高出 10%；值得注意的是，鉴于我们的分析完全彩票计划

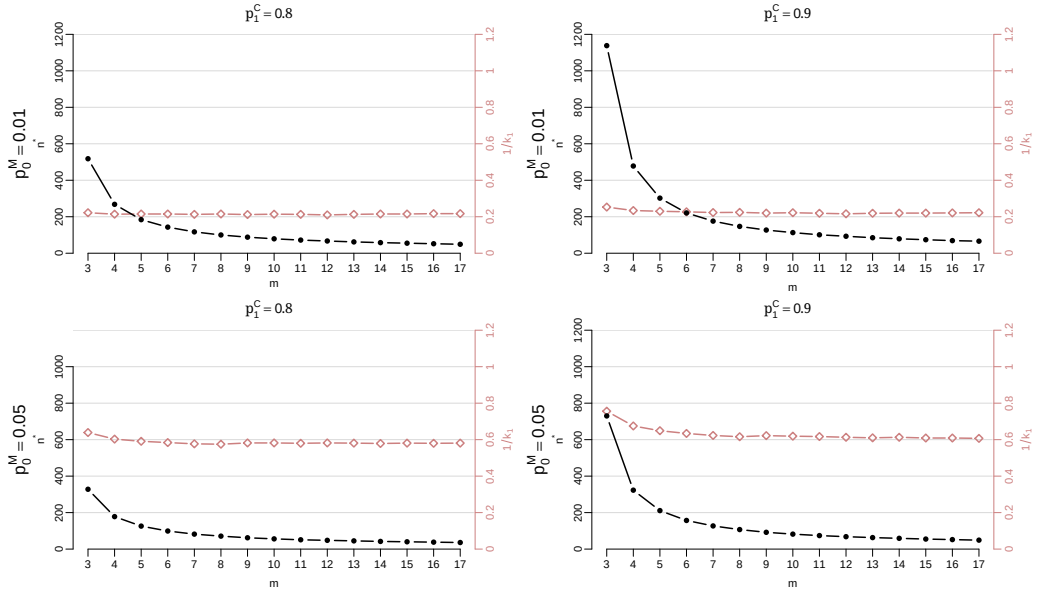


图5： $p_1^C \in \{0.8, 0.9\}$和 $p_0^M \in \{0.01, 0.05\}$对应的配对值（m、n^*）（实心圆），以及条件方法中的相应阈值 $1/k_1$（空心菱形）。

通过对 γ 采用完整的分布，纳入了参数 γ 的不确定性。通过使设计先验的方差趋近于零，我们可以得到与条件功效方法下的样本量结果类似的结果。

3.3 无条件方法的样本量确定

3.3.1 样本量确定的无条件标准

在 SSD 的无条件方法中，我们考虑表 1 中定义的总体概率p^C 和 p^M 。因此，我们通过将条件概率 p_1^C和 p_0^M替换为相应的总体概率，修改设计准则 (7)，从而得到 n^* 。

分别对应的总体概率 p^C 和 p^M。使用第 3.2.1 小节的记号，给定（$1-\beta$）、α 和研究中心数 m，按照 (7)，最优样本量n^* 定义如下:

$$n^* = \min\{n \in \mathbb{N} : p^C \geq (1-\beta) \ \ and \ \ p^M = \alpha\}, \tag{8}$$

其中， $\mathbb{N}$ 是自然数集合，为简单起见，我们设定 $p_0^M = p_1^M = \alpha$。注意，现在不同于条件方法，需要模型先验概率。对于表 3 中的结果，我们另外固定了$\pi_0 = \pi_1 = 0.5$。应用与条件方法中所用的类似计算，我们得到 $1/k_1$，此外，在约束 $p_1^M = \alpha$ 下，我们推导出 k_0，其为在给定 $\mathcal{M}_1$ 条件下 BF_{01} 的分布的（$1-\alpha$）$-$分位数。再次， n^* 的有效计算是通过 通过 假位法方法进行的，如条件方法中一样。

3.3.2 无条件方法的结果

在 $p^C \in \{0.80, 0.90\}$ 和 $p^M \in \{0.01, 0.05\}$ 的组合下，我们采用无条件方法，同时保持第 3.2.2 小节中报告的其他设置不变。结果列于表 3，并以图形形式呈现在图 6 中。

与条件方法类似，所需样本量取决于预先设定的概率 p^C 和 p^M，而 m 和 n 再次呈反比关系。就阈值k_0 和 $1/k_1$所隐含的证据尺度而言，两种模型的范围均为从“轶事性”到“中等”。由于无条件方法的性质更为严格，对于每个 m，与条件方法相比，无条件方法需要更大的样本量才能达到所需的正确证据概率，不过随着 m 增大，差异会逐渐缩小。看来，为无条件方法所需的更大样本量是值得的，因为我们不仅能控制

表 3：无条件方法中 $p^C \in \{0.80,\ 0.90\}$ 和 $p^M \in \{0.01, 0.05\}$ 对应的数对集合 $(n^*,\ m)$，以及决策阈值 $1/k_1$ 和 k_0 。

$p^C = 0.8$						$p^C = 0.9$					
$p^M = 0.01$			$p^M = 0.05$			$p^M = 0.01$			$p^M = 0.05$		
$(n^*,\ m)$	$1/k_1$	k_0	$(n^*,\ m)$	$1/k_1$	k_0	$(n^*,\ m)$	$1/k_1$	k_0	$(n^*,\ m)$	$1/k_1$	k_0
(2689, 3)	0.321	4.090	(611, 3)	0.724	2.049	(4596, 3)	0.386	3.301	(1018, 3)	0.827	1.518
(735, 4)	0.258	2.891	(266, 4)	0.647	1.761	(1130, 4)	0.292	2.257	(406, 4)	0.714	1.289
(404, 5)	0.244	2.525	(171, 5)	0.622	1.632	(590, 5)	0.269	1.930	(250, 5)	0.675	1.190
(265, 6)	0.234	2.273	(127, 6)	0.609	1.542	(374, 6)	0.254	1.726	(180, 6)	0.654	1.121
(197, 7)	0.227	2.128	(102, 7)	0.597	1.506	(274, 7)	0.244	1.603	(142, 7)	0.637	1.103
(158, 8)	0.227	2.015	(85, 8)	0.590	1.463	(214, 8)	0.241	1.537	(117, 8)	0.627	1.066
(133, 9)	0.221	1.972	(74, 9)	0.598	1.419	(179, 9)	0.234	1.504	(100, 9)	0.632	1.055
(117, 10)	0.223	1.918	(66, 10)	0.597	1.411	(155, 10)	0.234	1.445	(89, 10)	0.629	1.048
(103, 11)	0.220	1.886	(59, 11)	0.592	1.413	(136, 11)	0.231	1.421	(80, 11)	0.626	1.005
(91, 12)	0.216	1.819	(55, 12)	0.593	1.364	(119, 12)	0.225	1.370	(73, 12)	0.622	1.019
(83, 13)	0.219	1.762	(51, 13)	0.591	1.361	(108, 13)	0.228	1.351	(67, 13)	0.618	1.023
(78, 14)	0.220	1.764	(47, 14)	0.588	1.381	(101, 14)	0.229	1.337	(63, 14)	0.621	0.972
(72, 15)	0.219	1.731	(44, 15)	0.587	1.364	(92, 15)	0.227	1.338	(58, 15)	0.615	1.000
(68, 16)	0.221	1.718	(42, 16)	0.586	1.360	(87, 16)	0.229	1.293	(55, 16)	0.613	1.007
(64, 17)	0.222	1.691	(40, 17)	0.587	1.358	(82, 17)	0.230	1.275	(53, 17)	0.618	0.960

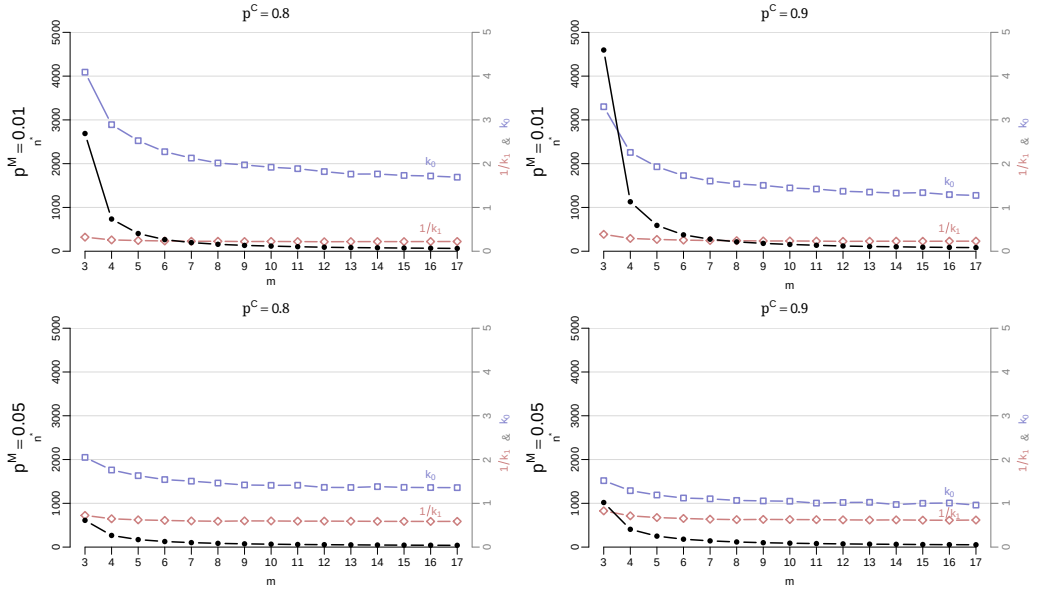


图 6：无条件方法下的 $p^C \in \{0.8, 0.9\}$ 和 $p^M \in \{0.01, 0.05\}$ 对应的数对 (m, n^*)（实心圆），以及相应的阈值 $1/k_1$ 和 k_0（分别为空心菱形和空心方块）。

每个模型下出现误导性证据的概率，但我们也对整体正确证据的概率有一个下界。然而，后者也间接地给出了每个模型下正确证据概率的界限，尤其是在 $(1-\beta)$ 较高且两个模型的概率不处于范围 (0，1) 的极端时。

与设计先验的变化相比，SSD 对设计先验变化的敏感性似乎比条件方法更为显著。具体而言，相对于基准情形$\mu_\gamma = 0.2$，设定为 $\mu_\gamma = 0.1$ 时，平均所需样本量是基准情形的六倍；而设定为 $\mu_\gamma = 0.3$ 时，样本量则减少了一半以上（约为基准情形的 40%）。详细结果见补充材料 B。

4 讨论

本文研究了多项重复研究，重点关注各研究中心效应量的变异（异质性）。具体而言，我们比较了两个模型：一个不考虑异质性（$\mathcal{M}_0$），另一个则纳入异质性（$\mathcal{M}_1$）。在此框架下，我们开发了一种贝叶斯样本量确定（SSD）程序，可用于获得有力证据。对于所考虑的两个模型，证据以贝叶斯因子（BF）表示，该因子基于恰当定义的分析先验推导而来。我们的设计准则通过条件方法和无条件方法加以规定。前者旨在以较高概率正确获得支持 $\mathcal{M}_1$ （存在异质性）的证据，同时确保基于 BF 的第一类错误率保持较低。无条件方法则旨在总体上（即对两个模型取平均）以较高概率获得正确证据，同时将误导性证据的概率总体上保持在较低水平。我们通过 BF 的先验预测分布来评估这一准则；该分布基于对设计先验（与分析先验相互独立）的设定而推导得出。

本文提出的贝叶斯方法，为重复研究中的 SSD 提供了一种灵活的替代方案，可替代基于频率学派的设计。我们的方法一大特点是在分析阶段和设计阶段都通过先验分布纳入不确定性，从而显著拓展了传统功效分析中以固定参数值为条件的标准做法。更广泛地说，我们可以利用贝叶斯因子进行证据评估所固有的优势（如 Wagenmakers 等（2016）所述），尤其是评估异质性存在与否的后验概率。

为简化阐述并聚焦于方法中最相关的方面，我们的设计未纳入成本考量。不过，这些考量也可以相当直接地纳入我们的框架。将总成本 C 表示为每名受试者成本c_1和每个实验室成本 c_2，只需在所有满足设计准则（7）或（8）的配对（n^*，m）中（见表2或表 3），选出使总成本 C最小的配对（n'，m'）。

贝叶斯推断及其设计中的一个重要问题，是结果对先验设定的敏感性。我们针对相对异质性设计先验进行了敏感性分析。正如预期，所需样本量与相对异质性参数的位置呈反比；详情及更多结果见补充材料。

尽管经过了周密规划，实际开展实验后，结果仍可能只能提供未定证据（见表 1），因此两个假设都得不到支持。在此阶段，一个自然的选择是规划后续设计，直至获得支持任一方向的有力证据，即支持 $\mathcal{M}_0$ 或 $\mathcal{M}_1$。这与 Sch¨onbrodt 等人（2017）所述的序贯贝叶斯因子思路一致，可采用开放式模式，也可采用最大样本量模式。

软件

用于复现论文中模拟结果的 R 代码，以及用于在不同于本研究所采用的设置下实现我们方法的 R 函数，可在 https://github.com/bourazaskonstantinos/Bayesian- SSD 获取。

致谢

本研究部分获得了加州大学圣克鲁兹分校（D1 研究资助）的支持。

利益披露声明

作者声明不存在需要披露的利益冲突。

参考文献

Baker, M. (2016年) 可重复性危机。《自然》, **533**, 353–366。

Bayarri, M. and Mayoral, A. (2002a) 贝叶斯分析与效应量比较的设计。《统计规划与推断杂志》, **103**, 225–243。

—（2002b）“成功”重复研究的贝叶斯设计。《美国统计学家》，**56**，207–214。

Bonett, D. G.（2012）重复与扩展研究。《心理科学当前方向》，**21**，409–412。

—（2021）重复研究的设计与分析。《组织研究方法》，**24**，513–529。

Burden, R. L.、Faires, J. D. 和 Burden, A. M.（2015）数值分析。圣智学习。

Dawid, A. P. (2011) 后验模型概率。载于《统计学哲学》（编者：P.S.Bandyopadhyay 和 M. Forster），607–630。爱思唯尔，阿姆斯特丹。

De Santis, F. (2004) 贝叶斯假设检验的统计证据与样本量确定。《统计规划与推断杂志》，**124**，121–144。Dienes, Z. (2014) 运用贝叶斯方法，充分利用不显著结果。*Frontiers in Psychology*，**5**，1–20。Etz, A. 和 Vandekerckhove, J. (2016) 关于可重复性项目的贝叶斯视角：心理学。*PloS one*，**11**，1–12。Fedorov, V. 和 Jones, B. (2005) 多中心试验设计。《医学研究统计方法》，**14**，205–248。Francis, G. (2012) 实验心理学中的发表偏倚与重复研究失败。*Psychonomic Bulletin & Review*，**19**，975–991。Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M. 和 Wagenmakers, E.-J. (2021) 贝叶斯模型平均元分析入门。《心理科学方法与实践进展》，**4**，1–19。Harden, M. 和 Friede, T. (2018) 多中心临床试验中的样本量计算。《*BMC*医学研究方法学》，**18**，1–10。Harms, C. (2019) 用于方差分析结果重复研究的贝叶斯因子。*The American Statistician*，**73**，327–339。Hedges, L. V. 和 Pigott, T. D. (2001) 元分析中统计检验的效力。*Psychological methods*，**6**，203–217。Hedges, L. V. 和 Schauer, J. M. (2019a) 要对重复研究进行明确检验，需要不止一项重复研究。《教育与行为统计学杂志》，**44**，543–570。

—（2019b）用于研究重复研究的统计分析：荟萃分析视角。*Psychological Methods*，**24**，557–570。—（2021）重复研究设计。《皇家统计学会期刊：*A*辑（社会统计学）》，**184**，868–886。Held, L.（2020）重复研究分析与设计的新标准。《皇家统计学会期刊：*A*辑（社会统计学）》，**183**，431–448。Held, L.、Matthews, R.、Ott, M. 和 Pawel, S.（2022）用于证据评估和研究综合的逆贝叶斯方法。*Research SynthesisMethods*，**13**，295–314。

Hensel, P. G.（2021）可重复性与可复现性危机：管理学与心理学和经济学的比较——文献系统综述。《欧洲管理学杂志》，**39**，577–594。

Hoijtink, H., Mulder, J., van Lissa, C.和 Gu, X. (2019) 使用贝叶斯因子检验假设教程。*Psychological methods*，**24**，539–556。

Hou, K., Xue, C. 和 Zhang, L.（2020）重复研究异常现象。《金融研究评论》，**33**，2019–2133。Ioannidis, J. P.（2005）为什么大多数已发表研究结果都是错误的。《公共科学图书馆医学》，**2**，696–701。Jeffreys, H.（1961）《概率论》（第*3*版）。牛津：牛津大学出版社。Johnson, V. E. 和 Rossell, D.（2010）论贝叶斯假设检验中非局部先验密度的使用。《皇家统计学会期刊：*B*辑》，**72**，143–170。

Kass, R. E. 和 Raftery, A. E.（1995）贝叶斯因子。美国统计学会会刊，90，773–795。Klein, S. B.（2014）近期的重复失败能告诉我们心理学的理论承诺些什么？理论与心理学，24，326–338。Ly, A. 和 Wagenmakers, E.-J.（2022）近零假设的贝叶斯因子。*TEST*。网址：https://doi.org/10.1007/s11749-022-00819-w。Marsman, M., Sch¨onbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A. 和 Wagenmakers, E.-J.（2017）从贝叶斯视角纵览“社会心理学重要结果的重复研究”。英国皇家学会开放科学，4，1–18。Mathur, M. B. 和 VanderWeele, T. J.（2020）多地点重复项目的新统计指标。《皇家统计学会期刊：*A*辑（社会统计学）》，183，1145–1166。Muradchanian, J., Hoekstra, R., Kiers, H. 和 van Ravenzwaaij, D.（2021）如何最好地量化重复成功？一项比较重复成功指标的模拟研究。英国皇家学会开放科学，8，1–16。O’Hagan, A., Stevens, J. W. 和 Campbell, M. J.（2005）临床试验设计中的把握度评估。《药物统计学：制药行业应用统计学期刊》，4，187–201。O’Hagan, A. 和 Stevens, J. W.（2001）对成本效益临床试验样本量的贝叶斯评估。《医疗决策》，21，219–230。Pashler, H. 和 Harris, C. R.（2012）可复现性危机是否被夸大了？对三个论点的考察。*Perspectives on Psychological Science*，7，531–536。

Pawel, S. 和 Held, L.（2020年）重复研究的概率预测。*PloS one*，**15**，1–23。Psarakis, S. 和 Panaretoes, J.（1990年）折叠 t 分布。*Communi-cations in Statistics-Theory and Methods*，**19**，2717–2734。R¨over, C.、Bender, R.、Dias, S.、Schmid, C. H.、Schmidli, H.、Sturtz, S.、Weber, S. 和 Friede, T.（2021年）关于贝叶斯随机效应荟萃分析中异质性参数的弱信息先验分布。*Research Synthesis Methods*，**12**，448–474。Sch¨onbrodt, F. D. 和 Wagenmakers, E.-J.（2018年）贝叶斯因子设计分析：为有力证据进行规划。*Psychonomic bulletin & review*，**25**，128–142。Sch¨onbrodt, F. D.、Wagenmakers, E.-J.、Zehetleitner, M. 和 Perugini, M.（2017年）使用贝叶斯因子进行序贯假设检验：高效检验均值差异。*Psychological Methods*，**22**，322–339。Simon, R.（1999年）主动对照临床试验的贝叶斯设计与分析。*Biometrics*，**55**，484–487。Simons, D. J.（2014年）直接重复的价值。*Perspectives on Psycho-logical Science*，**9**，76–80。Verhagen, J. 和 Wagenmakers, E.-J.（2014年）用于量化重复尝试结果的贝叶斯检验。*Journal of Experimental Psychology: Gen-eral*，**143**，1457–1475。Wagenmakers, E.-J.、Beek, T. F.、Rotteveel, M.、Gierholz, A.、Matzke, D.、Steingroever, H.、Ly, A.、Verhagen, J.、Selker, R.、Sasiadek, A. 等（2015年）再次拨动时间之手：一项纯验证性重复研究与贝叶斯分析。*Frontiers in Psychology*，**6**，1–6。

Wagenmakers, E.-J.、Morey, R. D. 和 Lee, M. D.（2016年）贝叶斯方法为务实研究者带来的益处。*Current Directions in Psychological Science*, **25**，169–176。

Wang, F. 和 Gelfand, A. E.（2002）基于模拟的方法：在给定模型下评估表现并区分模型的贝叶斯样本量确定。《统计科学》，193–208。

Wasserstein, R. L. 和 Lazar, N. A.（2016）美国统计学会关于p值的声明：背景、过程与目的。《美国统计学家》，**70**，129–133。

Weiss, R.（1997年）用于假设检验的贝叶斯样本量计算。《英国皇家统计学会会刊：D辑（统计学家）》，**46**，185–191。

Wong, V. C.、Anglin, K. 和 Steiner, P.M.（2021）因果复制研究的设计型方法。《预防科学》，**23**，1–16。

Zwaan, R. A.、Etz, A.、Lucas, R. E. 和 Donnellan, M. B.（2018年）让重复研究成为主流。《行为与脑科学》，**41**，1–61。

Zwanenburg, A.（2019）核医学中的影像组学：稳健性、可重复性、标准化，以及如何避免数据分析陷阱和重复危机。《欧洲核医学与分子影像杂志》，**46**，2638–2655。

附录

在本附录中，我们给出在 $\mathcal{M}_0$) 和 $\mathcal{M}_1$) 条件下边际数据分布的解析推导，然后给出贝叶斯因子。

考虑随机向量 $\boldsymbol{t_r}$ =(t_1、.....、t_m)，其中 t_j是位点 $j, j = 1$、.....、m 的效应量估计。在模型 $\mathcal{M}_0$下，数据分布为 $t_j|\mu$，$\sim N(\mu$，$\sigma^2/n)$；而在模型 $\mathcal{M}_1$下，数据分布为 $t_j|\mu$，$\gamma \sim N(\mu$，$\sigma^2(1/n + \gamma^2))$，且所有变量相互独立。令 $Q = n\cdot\sum_{j=1}^{m}(t_j - \bar{t})^2/\sigma^2$，其中 $\bar{t}$是 $\boldsymbol{t_r}$的样本均值，则在 $\mathcal{M}_0$ 下的边际数据分布为

$$
\begin{aligned}
m_0(\boldsymbol{t_r}) &= \int_{-\infty}^{+\infty} f(\boldsymbol{t_r}|\mu, \mathcal{M}_0)\pi(\mu)d\mu = \int_{-\infty}^{+\infty} \left(\frac{1}{2\pi\sigma^2/n}\right)^{m/2} \cdot \exp\left\{-\frac{\sum_{i=1}^{m}(t_i-\mu)^2}{2\sigma^2/n}\right\} d\mu \\
&= \left(\frac{2\pi\sigma^2}{n}\right)^{-m/2} \cdot \exp\left\{-\frac{\sum_{i=1}^{m}(t_i-\bar{t})^2}{2\sigma^2/n}\right\} \int_{-\infty}^{+\infty} \exp\left\{-\frac{\sum_{i=1}^{m}(\bar{t}-\mu)^2}{2\sigma^2/n}\right\} d\mu \\
&= \left(\frac{2\pi\sigma^2}{n}\right)^{-m/2} \cdot \exp\left\{-\frac{\sum_{i=1}^{m}(t_i-\bar{t})^2}{2\sigma^2/n}\right\} \cdot \left(\frac{2\pi\sigma^2}{n\cdot m}\right)^{1/2} \\
&= \left(\frac{1}{m}\right)^{1/2} \cdot \left(\frac{2\pi\sigma^2}{n}\right)^{(1-m)/2} \cdot \exp\left\{-\frac{Q}{2}\right\}.
\end{aligned}
$$

.

另一方面，在 $\mathcal{M}_1$ 下的边际数据分布为

$$
\begin{aligned}
m_1(\boldsymbol{t_r}) &= \int_0^{+\infty}\int_{-\infty}^{+\infty} f(\boldsymbol{t_r}|\mu,\gamma,\mathcal{M}_1)h_a(\gamma)\pi(\mu)d\mu d\gamma \\
&= \int_0^{+\infty}\int_{-\infty}^{+\infty} \left(2\pi\sigma^2\left(\frac{1}{n}+\gamma^2\right)\right)^{-m/2}\cdot\exp\left\{-\frac{n\sum_{i=1}^m(t_i-\mu)^2}{2\sigma^2\left(1+n\gamma^2\right)}\right\}h_a(\gamma)\pi(\mu)d\mu d\gamma \\
&= \int_0^{+\infty} \left(2\pi\sigma^2\left(\frac{1}{n}+\gamma^2\right)\right)^{-m/2}\cdot\exp\left\{-\frac{n\sum_{i=1}^m(t_i-\bar{t})^2}{2\sigma^2\left(1+n\gamma^2\right)}\right\}h_a(\gamma)\times \\
&\times\int_{-\infty}^{+\infty}\exp\left\{-\frac{n\sum_{i=1}^m(\bar{t}-\mu)^2}{2\sigma^2\left(1+n\gamma^2\right)}\right\}d\mu d\gamma \\
&= \int_0^{+\infty} \left(2\pi\sigma^2\left(\frac{1}{n}+\gamma^2\right)\right)^{-m/2}\cdot\exp\left\{-\frac{n\sum_{i=1}^m(t_i-\bar{t})^2}{2\sigma^2\left(1+n\gamma^2\right)}\right\} \\
&\times\left(\frac{2\pi\sigma^2}{m}\left(\frac{1}{n}+\gamma^2\right)\right)^{1/2}h_a(\gamma)d\gamma \\
&= \left(\frac{1}{m}\right)^{1/2}\cdot\left(2\pi\sigma^2\right)^{(1-m)/2}\int_0^{+\infty}\left(\frac{1}{n}+\gamma^2\right)^{(1-m)/2}\exp\left\{-\frac{Q}{2}\cdot\left(1+n\gamma^2\right)^{-1}\right\}h_a(\gamma)d\gamma.
\end{aligned}
$$

由此得到的贝叶斯因子为:

$$
BF_{01}(\boldsymbol{t_r}) = \frac{m_0(\boldsymbol{t_r})}{m_1(\boldsymbol{t_r})} = \frac{n^{(m-1)/2}\cdot\exp\left\{-\frac{Q}{2}\right\}}{\int_0^{+\infty}\left(\frac{1}{n}+\gamma^2\right)^{(1-m)/2}\exp\left\{-\frac{Q}{2}\cdot(1+n\gamma^2)^{-1}\right\}h_a(\gamma)d\gamma}.
$$

补充材料

A. 条件方法的设计先验敏感性分析

我们针对第 3.2 小节提出的条件方法的设计先验进行了敏感性分析。具体而言，我们考虑另外两个设计先验，它们与正文中的先验仅在位置超参数 μ_γ 上有所不同。在第一个分布中，我们设定 $\mu_\gamma = 0.1$，得到的 95% 可信区间为 $[0.05, 0.15]$。从实践角度来看，这个新的先验集中在 μ_γ 的较低取值上，代表一种较为敏感的设定，在这种设定下，我们希望检测到较小的相对异质性。另一方面，第二个设计先验集中在 μ_γ 的较高取值上，代表对异质性更为宽容的水平。具体而言，我们设定 $\mu_\gamma = 0.3$，由此得到的 95% 可信区间为 $[0.25, 0.35]$。图 7 展示了这两个额外先验的密度图，以及正文中的设计先验（$\mu_\gamma = 0.2$）和分析先验的密度图。

表 1 和表 2 给出了敏感性分析的结果，相应的图表见图 8 和图 9。正如预期，当相对异质性 γ 的设计先验向较低值偏移时，需要更大的样本量；而当分布向 γ 的较高值偏移时，情况则相反。另一方面，支持 $\mathcal{M}_1$ 的证据尺度上的阈值 $1/k_1$ 在不同设计先验下并未改变，即 $p_0^M = 0.05$ 对应“轶事性”，$p_0^M = 0.01$对应“中等”。

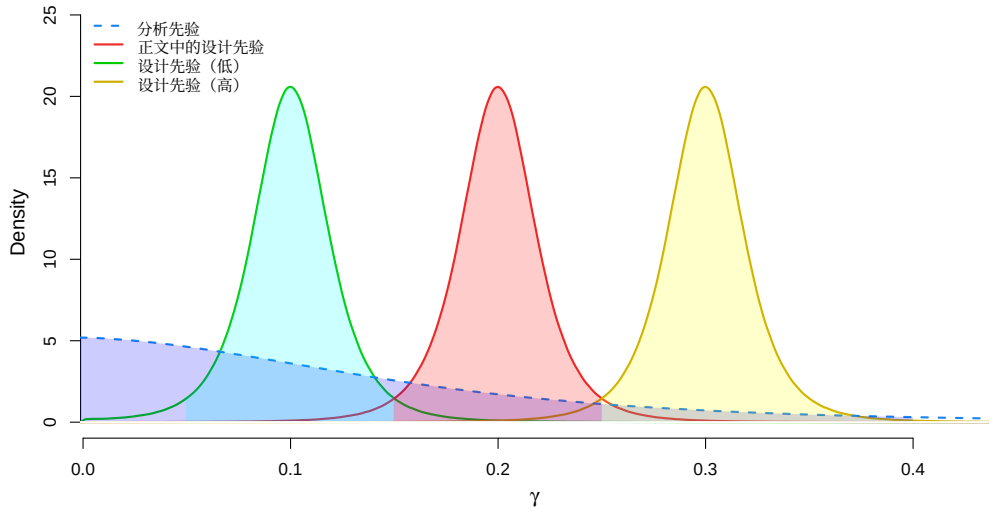


图 7：分析先验 $h_a(\gamma)$（虚线）和三个设计 先验 $h_d(\gamma)$（实线），以及它们的 95% 可信区间（高亮区域）。

表 4: $p_1^C \in \{0.8, 0.9\}$ 和 $p_0^M \in \{0.01, 0.05\}$的数对集合 (n^*, m)，以及条件方法对应的阈值 $1/k_1$。所使用的设计先验为折叠-t($\nu = 4,\ \mu_\gamma = 0.1,\ \sigma_\gamma = 1/55$)。

$p_1^C = 0.8$				$p_1^C = 0.9$			
$p_0^M = 0.01$		$p_0^M = 0.05$		$p_0^M = 0.01$		$p_0^M = 0.05$	
$(n^*,\ m)$	$1/k_1$	$(n^*,\ m)$	$1/k_1$	$(n^*,\ m)$	$1/k_1$	$(n^*,\ m)$	$1/k_1$
(2314, 3)	0.307	(1451, 3)	0.920	(5294, 3)	0.407	(3382, 3)	1.236
(1192, 4)	0.297	(785, 4)	0.869	(2270, 4)	0.373	(1521, 4)	1.105
(824, 5)	0.298	(557, 5)	0.855	(1443, 5)	0.365	(996, 5)	1.062
(639, 6)	0.298	(439, 6)	0.850	(1053, 6)	0.357	(739, 6)	1.034
(519, 7)	0.293	(360, 7)	0.837	(841, 7)	0.350	(597, 7)	1.012
(445, 8)	0.297	(312, 8)	0.834	(699, 8)	0.351	(501, 8)	0.998
(394, 9)	0.292	(274, 9)	0.848	(608, 9)	0.342	(432, 9)	1.008
(353, 10)	0.295	(246, 10)	0.845	(540, 10)	0.344	(387, 10)	1.006
(320, 11)	0.292	(225, 11)	0.844	(487, 11)	0.340	(350, 11)	0.999
(296, 12)	0.288	(208, 12)	0.842	(447, 12)	0.335	(321, 12)	0.994
(275, 13)	0.293	(195, 13)	0.840	(410, 13)	0.339	(297, 13)	0.988
(259, 14)	0.295	(184, 14)	0.845	(380, 14)	0.340	(277, 14)	0.987
(243, 15)	0.294	(173, 15)	0.841	(356, 15)	0.338	(260, 15)	0.984
(230, 16)	0.297	(164, 16)	0.839	(338, 16)	0.342	(247, 16)	0.984
(218, 17)	0.297	(156, 17)	0.841	(318, 17)	0.341	(234, 17)	0.985

表5： $p_1^C \in \{0.8, 0.9\}$ 和 $p_0^M \in \{0.01, 0.05\}$的数对集合（n^*， m），以及条件方法对应的阈值 $1/k_1$。所用的设计先验为折叠-t($\nu = 4$， $\mu_\gamma = 0.3$， $\sigma_\gamma = 1/55$)。

$p_1^C = 0.8$				$p_1^C = 0.9$			
$p_0^M = 0.01$		$p_0^M = 0.05$		$p_0^M = 0.01$		$p_0^M = 0.05$	
(n^*, m)	$1/k_1$	(n^*, m)	$1/k_1$	(n^*, m)	$1/k_1$	(n^*, m)	$1/k_1$
(226，3)	0.209	(143，3)	0.583	(490, 3)	0.218	(315，3)	0.635
(116, 4)	0.211	(77, 4)	0.561	(206、4)	0.210	(139, 4)	0.584
(80，5)	0.216	(55, 5)	0.556	(130，5)	0.211	(91，5)	0.569
(62，6)	0.218	(43, 6)	0.554	(96，6)	0.212	(68，6)	0.562
(51，7)	0.219	(36, 7)	0.551	(76，7)	0.211	(55，7)	0.557
(44，8)	0.222	(31, 8)	0.551	(63, 8)	0.215	(46，8)	0.553
(39，9)	0.220	(27，9)	0.560	(55，9)	0.212	(40，9)	0.562
(35, 10)	0.223	(25，10)	0.560	(48，10)	0.215	(36, 10)	0.562
(32, 11)	0.222	(23，11)	0.560	(44, 11)	0.214	(32, 11)	0.560
(29, 12)	0.222	(21，12)	0.561	(40，12)	0.212	(29，12)	0.559
(27，13)	0.226	(20, 13)	0.559	(37, 13)	0.216	(27，13)	0.558
(25，14)	0.229	(18，14)	0.563	(34, 14)	0.218	(25，14)	0.559
(24, 15)	0.228	(17，15)	0.563	(32，15)	0.218	(24，15)	0.560
(23，16)	0.228	(17，16)	0.560	(30，16)	0.210	(23, 16)	0.559
(22，17)	0.229	(16, 17)	0.562	(29, 17)	0.220	(22, 17)	0.561

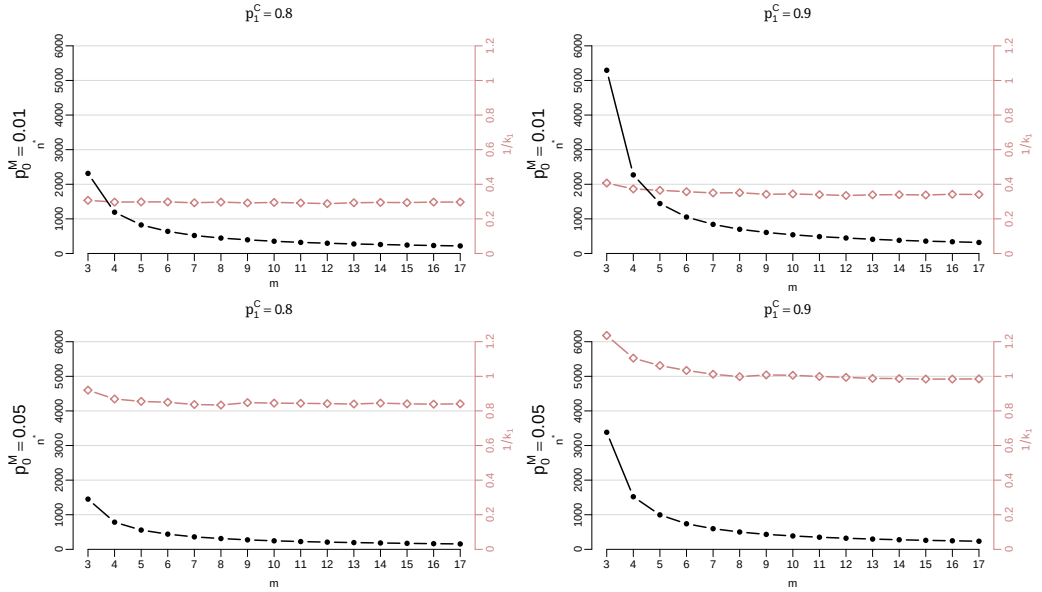


图 8: $p_1^C \in \{0.8, 0.9\}$ 和 $p_0^M \in \{0.01, 0.05\}$的配对（m, n^*）（实心圆），以及条件方法对应的阈值 $1/k_1$（空心菱形）。所使用的折叠-t($\nu = 4$, μ_γ =0。1, $\sigma_\gamma = 1/55$)设计先验。

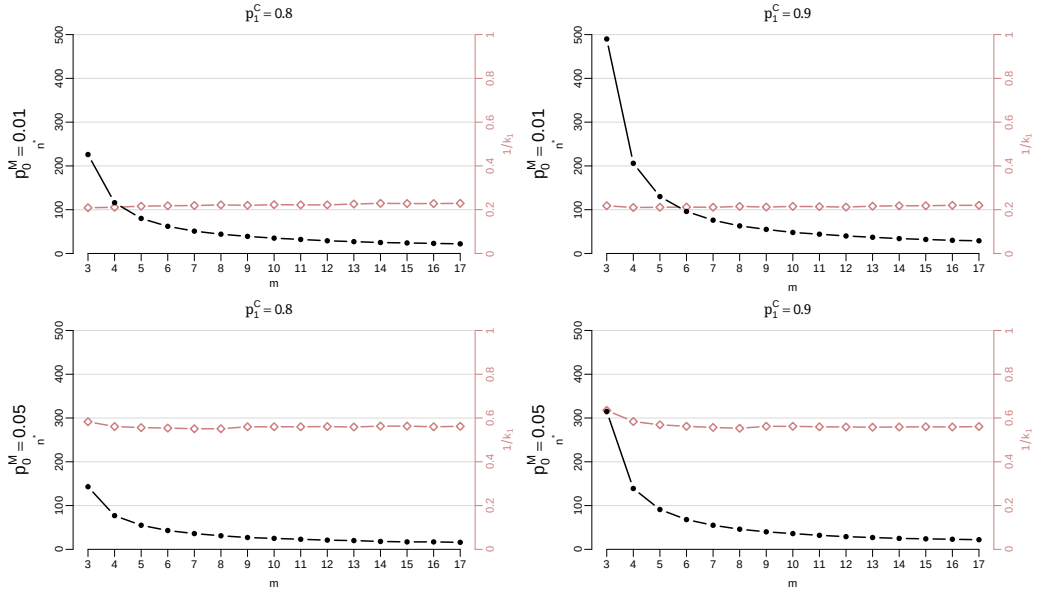


图 9：条件方法中，$p_1^C \in \{0.8, 0.9\}$ 和 $p_0^M \in \{0.01, 0.05\}$对应的配对值 (m, n^*)（实心圆）以及相应的阈值 $1/k_1$（空心菱形）。采用的设计先验为折叠-$t(\nu = 4,\ \mu_\gamma = 0.3,\ \sigma_\gamma = 1/55)$。

B. 无条件方法的设计先验敏感性分析

此处，我们展示了无条件方法在另外两种设计先验下的表现，即补充材料 A 中介绍的先验，其相对异质性分别处于 5%-15% 和 25%-35% 的范围内。模拟结果列于表 3 和表 4，图形表示见图 10 和图 11。与正文中的假设相同，我们假设 $p_0^M = p_1^m = \alpha$ 和 $\pi_0 = \pi_1 = 0.5$。

关于设计先验中 μ_γ 变化所产生影响的讨论，与补充材料 A 中的讨论类似。值得一提的是，对于无条件方法，在某些情况下，证据尺度会失去其原有的解释。具体而言，当设计先验非常接近 0 或远离 0 时，分别极难获得支持 $\mathcal{M}_1$ 或 $\mathcal{M}_0$ 的有力证据。因此，我们可能会得到分别大于或小于 1 的 $1/k_1$ 或 k_0 值。

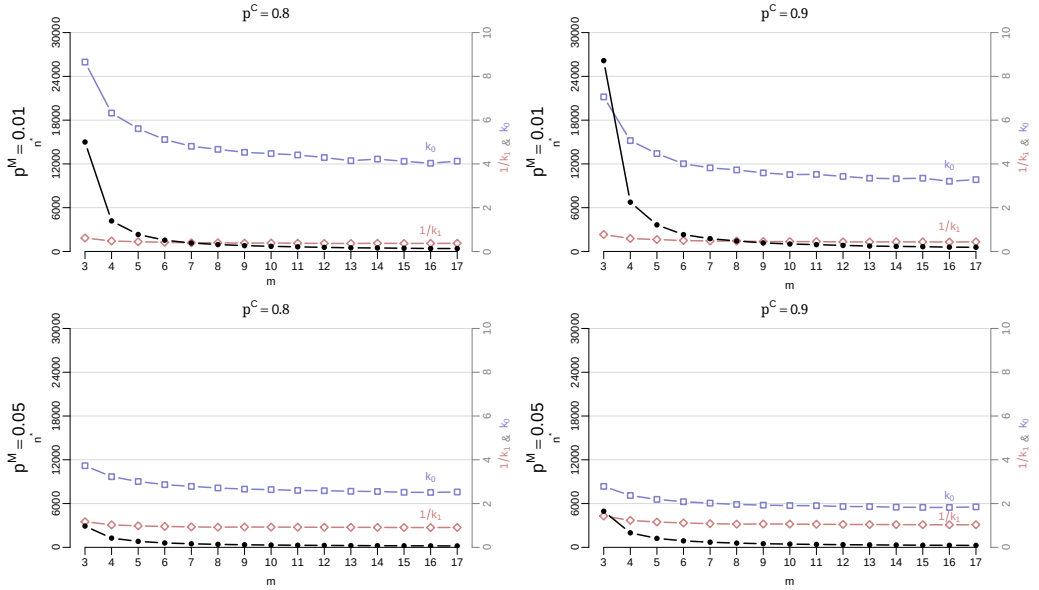


图**10**：无条件方法下的配对值（m，n^*）（实心圆），对应于 $p^C \in \{0.8, 0.9\}$ 和 $p^M \in \{0.01, 0.05\}$；以及相应的阈值 $1/k_1$ 和 k_0（分别为空心菱形和空心正方形）。所用的设计先验为折叠-*t*($\nu = 4$，$\mu_\gamma = 0.1$，$\sigma_\gamma = 1/55$)。

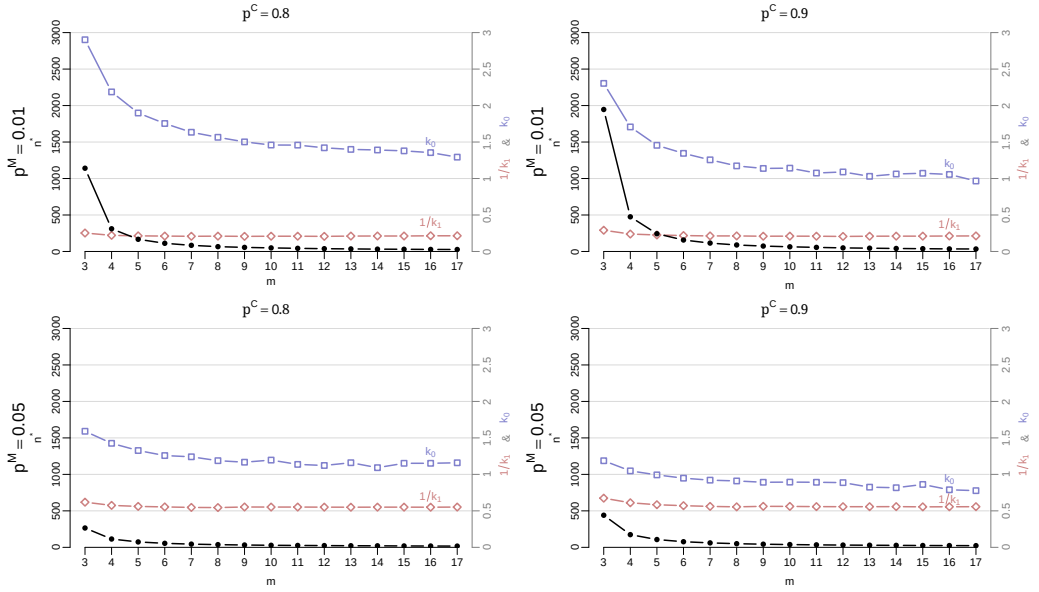


图**11**：无条件方法中，$p^C \in \{0.8, 0.9\}$ 和 $p^M \in \{0.01, 0.05\}$对应的配对（m，n^*）（实心圆），以及相应的阈值 $1/k_1$ 和 k_0（分别为空心菱形和空心方形）。使用的设计先验为折叠-t($\nu = 4$，$\mu_\gamma = 0.3$，$\sigma_\gamma = 1/55$)。

表 6：无条件方法的数对集合（n^*，m）及其决策阈值$1/k_1$ 和 k_0。所用的设计先验为折叠-t($\nu = 4$，$\mu_\gamma = 0.1$，$\sigma_\gamma = 1/55$)。

$p^C = 0.8$						$p^C = 0.9$					
$p^M = 0.01$			$p^M = 0.05$			$p^M = 0.01$			$p^M = 0.05$		
$(n^*,\ m)$	$1/k_1$	k_0	$(n^*,\ m)$	$1/k_1$	k_0	$(n^*,\ m)$	$1/k_1$	k_0	$(n^*,\ m)$	$1/k_1$	k_0
(15001, 3)	0.614	8.654	(2915, 3)	1.170	3.730	(26156, 3)	0.777	7.061	(4944, 3)	1.431	2.785
(4188, 4)	0.479	6.324	(1263, 4)	1.028	3.228	(6756, 4)	0.591	5.066	(1991, 4)	1.231	2.368
(2326, 5)	0.445	5.610	(813, 5)	0.981	3.010	(3648, 5)	0.544	4.474	(1225, 5)	1.156	2.191
(1545, 6)	0.418	5.115	(601, 6)	0.954	2.865	(2311, 6)	0.500	4.008	(886, 6)	1.114	2.084
(1159, 7)	0.399	4.806	(486, 7)	0.934	2.786	(1757, 7)	0.480	3.820	(702, 7)	1.082	2.019
(950, 8)	0.397	4.666	(408, 8)	0.920	2.712	(1438, 8)	0.477	3.726	(579, 8)	1.059	1.961
(799, 9)	0.382	5.530	(350, 9)	0.928	2.661	(1175, 9)	0.453	3.591	(498, 9)	1.068	1.928
(706, 10)	0.384	4.470	(314, 10)	0.926	2.636	(1023, 10)	0.452	3.513	(443, 10)	1.063	1.909
(634, 11)	0.378	4.410	(284, 11)	0.920	2.601	(937, 11)	0.449	3.523	(400, 11)	1.054	1.902
(564, 12)	0.367	4.292	(262, 12)	0.916	2.588	(832, 12)	0.435	3.428	(364, 12)	1.047	1.865
(500, 13)	0.367	4.148	(243, 13)	0.914	2.562	(745, 13)	0.436	3.347	(337, 13)	1.042	1.861
(481, 14)	0.373	4.221	(226, 14)	0.911	2.550	(695, 14)	0.438	3.324	(313, 14)	1.039	1.834
(440, 15)	0.368	4.119	(211, 15)	0.906	2.516	(657, 15)	0.437	3.347	(291, 15)	1.032	1.825
(405, 16)	0.368	4.031	(200, 16)	0.904	2.509	(592, 16)	0.432	3.208	(276, 16)	1.030	1.822
(401, 17)	0.375	4.126	(192, 17)	0.909	2.529	(578, 17)	0.438	3.281	(264, 17)	1.033	1.848

表 7：无条件方法的数对集合（n^*, m），以及决策阈值$1/k_1$ 和 k_0。所使用的设计先验为折叠-t($\nu = 4,\ \mu_\gamma = 0.3,\ \sigma_\gamma = 1/55$)。

$p^C = 0.8$						$p^C = 0.9$					
$p^M = 0.01$			$p^M = 0.05$			$p^M = 0.01$			$p^M = 0.05$		
$(n^*,\ m)$	$1/k_1$	k_0	$(n^*,\ m)$	$1/k_1$	k_0	$(n^*,\ m)$	$1/k_1$	k_0	$(n^*,\ m)$	$1/k_1$	k_0
(1141, 3)	0.253	2.902	(265, 3)	0.619	1.591	(1946, 3)	0.291	2.340	(440, 3)	0.674	1.187
(311, 4)	0.221	2.187	(114, 4)	0.575	1.410	(475, 4)	0.239	1.707	(174, 4)	0.611	1.048
(167, 5)	0.214	1.898	(74, 5)	0.561	1.327	(243, 5)	0.226	1.455	(107, 5)	0.584	0.993
(112, 6)	0.211	1.754	(55, 6)	0.554	1.258	(157, 6)	0.218	1.345	(77, 6)	0.571	0.948
(84, 7)	0.208	1.634	(44, 7)	0.547	1.241	(115, 7)	0.213	1.256	(61, 7)	0.561	0.921
(66, 8)	0.210	1.565	(37, 8)	0.545	1.188	(89, 8)	0.213	1.173	(50, 8)	0.555	0.910
(56, 9)	0.207	1.502	(32, 9)	0.553	1.167	(74, 9)	0.209	1.138	(43, 9)	0.562	0.892
(49, 10)	0.209	1.459	(28, 10)	0.552	1.196	(64, 10)	0.210	1.142	(38, 10)	0.561	0.894
(43, 11)	0.209	1.458	(26, 11)	0.552	1.137	(56, 11)	0.209	1.075	(34, 11)	0.558	0.892
(38, 12)	0.207	1.421	(24, 12)	0.551	1.122	(49, 12)	0.206	1.090	(31, 12)	0.557	0.887
(35, 13)	0.211	1.399	(22, 13)	0.550	1.160	(45, 13)	0.209	1.030	(29, 13)	0.557	0.825
(32, 14)	0.212	1.391	(21, 14)	0.551	1.092	(41, 14)	0.210	1.062	(27, 14)	0.558	0.817
(30, 15)	0.212	1.379	(19, 15)	0.551	1.152	(38, 15)	0.210	1.072	(25, 15)	0.555	0.862
(28, 16)	0.215	1.355	(18, 16)	0.550	1.152	(35, 16)	0.212	1.056	(24, 16)	0.556	0.789
(27, 17)	0.215	1.293	(17, 17)	0.552	1.159	(34, 17)	0.213	0.966	(23, 17)	0.557	0.777