标题：

希腊字母 π

副标题：

贝叶斯解释、p值的可重复性及在以下模型中的重复验证考量

广义线性模型

作者：

Christos Argyropoulos*‡ Andy P Grieve¶

*内科学系，新墨西哥大学医学院，MSC10 5550

1 美国新墨西哥州阿尔伯克基，新墨西哥州 87131，美国

¶ UCB Celltech 统计创新卓越中心

‡ 通讯作者：cargyropoulos@salud.unm.edu

摘要

基于 p值的显著性检验被认为是科学研究可重复性危机的成因之一，有人提出应弃用 p值，转而采用贝叶斯分析。p值的支持者则反驳说，问题出在使用不当和解释错误，而非 p值本身。自 p值正式引入统计实践以来，关于其作用的类似争论大约每隔 10 到 15 年就会出现一次。对 p值的解释屡屡失败，却仍持续使用，这种看似矛盾的现象表明，计算这些数值具有推断价值。在本文中，我们提出对广义线性模型中计算并报告为 p值的数值赋予一种激进的贝叶斯解释。我们证明，在用于论证非贝叶斯分析的数学假设下，前者与采用局部均匀先验的贝叶斯后验分布一致。接着，我们提出一个用于对后验尾部概率（π-值）进行阈值化的决策分析框架；对于任何给定的贝叶斯分析，该框架得出的数值都与非贝叶斯 p值完全相同。π值没有争议，但作为治疗效应的后验汇总，其意义相对有限。随后，我们提出一种基于预测概率的论证，以探讨 p 值和 π-值的随机变异（重复概率），从而说明仅依据 p值的统计显著性来判断研究发现能否重复验证，存在严重缺陷。

目录

1. 引言 4

2. 广义线性模型的似然方法 8

2.1. 广义线性模型治疗效应的估计 8 2.2. 尺度参数的估计 11 2.3. 广义线性模型中的 p值与零假设显著性检验 13

3. 广义线性模型的贝叶斯方法 16

3.1. 治疗效应的有界均匀先验 16 3.2. 广义线性模型的尺度参数先验分布 25 3.3. 贝叶斯分析与似然主义分析之间的数值等价 26 3.4. 魔法公式、非二次似然与局部均匀先验 33

4. Π 值 , T贝叶斯对应物 p 值 37

4.1. THE Π-值量化方向性协变量效应的幅度 37 4.2. TP 与 Π-值的数值关系 38 4.3. A 对 Π-值进行阈值化的决策分析框架 40

5. 重复验证 O F 科学发现与P值的可重复性 48

5.1. 用于重复验证科学发现的预测概率框架 49 5.2. P 值和 Π– 值的可重复性较差 53

6. 可重复 性：SGLT2i肾脏结局临床试验 57

7. 旨在实现重复验证的贝叶斯／似然主义综合分析 66

附录：在实数集中编写可重复性分析程序 82

A1. 可重复性计算 82 A2. 计算结果绘图 90

参考文献 94

主题索引 103

1. 引言

近期文献将 p 值显著性检验视为科学研究可重复性危机的罪魁祸首（Siegfried 2010；Colquhoun 2014, 2017；Wasserstein 等，2019；Gibson 2021）。有人提出贝叶斯方法（Wagenmakers 2007；Morey 等，2016a, b），认为这或许能弥补传统假设检验的不足。我们认为，贝叶斯主义者应当警惕一种魔法般的思维：不加区别地将贝叶斯方法奉为拯救科学免于不可重复性的良方。此外，针对许多与生物医学相关模型中 p 值的普遍批评，若不加以恰当限定，可能并不恰当，且会适得其反，不利于实现科学事业的可重复性目标。

本文阐述了几个彼此相关的观点，这些观点涉及 p 值在可重复性危机中的作用。首先，我们证明，对于常用的广义线性模型（GLM），非贝叶斯分析与采用均匀分布及其他“最少信息”局部 均匀分布先验分布的贝叶斯分析在数值上等价。这类先验分布（以及其他更具数学复杂性的选择）用于客观贝叶斯分析。客观贝叶斯分析可视为一系列从数据中学习的方法：当基于主观先验分布的分析站不住脚（Berger 2006），或会受到科学群体质疑时，可以采用这些方法。这种数值等价使我们能够对广义线性模型的非贝叶斯分析赋予贝叶斯解释。其次，可以为 p 值的计算及阈值化提供贝叶斯依据。本文将这些贝叶斯对应量称为 π 值；它们在数值上与 p 值完全相同，并为方向性统计决策（Kaiser 1960）以及以序数性主张表述的效应方向推断提供了方法（Chow 1988；Harris 1997；Nickerson 2000；Lakens 2021）。第三，我们提出一个用于评估特定研究结果可重复性的一般性预测概率框架。在最基础的层面上，也就是以与原研究完全相同的特征重复一项研究时，该框架使人们考察 p 值和 π 值的后验预测分布。第四，我们推导出其形式

，即 广义线性模型回归中 p 值的 “可重复性概率密度”（RPD），从而将二项比例、单样本和双样本比较以及多平行组设计的既有结果（Lambert 和 Hall 1982；Shao 和 Chow 2002；De Martini 2008；Boos 和 Stefanski 2011；De Capitani 和 De Martini 2016）扩展到广义线性模型回归。我们证明，对于这些广泛使用的广义线性模型，p 值和 π值存在显著变异，这与以往的观察结果一致（Senn 2002；Boos 和 Stefanski 2011）。因此，研究结果不一致（单项研究中出现“显著与不-显著”的情况）可能相当常见；这并非源于基础科学不可重复，而是 p 值和 π值本身具有噪声所致。

本专著的结构如下：第 2 节将完整阐述广义线性模型的非贝叶斯（基于似然）方法。第 3 节介绍在（局部）均匀先验下对广义线性模型进行贝叶斯分析。我们从三个不同角度论证在广义线性模型中使用这一先验的合理性：对有限效应进行推断；对于有界线性泛函关系的取值而言，它不提供信息；以及在未开展先验信息征询的情况下，将其作为研究者群体的参考先验。相应的后验分布与通常用于分析生物医学数据的广义线性模型的最大似然（最大似然）估计值具有完全相同的函数形式。其成立条件与似然函数渐近收敛到多元二次函数（这是最大似然估计量的基础）时的条件相同。在技术层面，我们表明，广义线性模型的各种非贝叶斯方法，例如扩展准似然、修正剖面似然或矩估计量，都可以解释为对完整贝叶斯分析进行近似计算。我们还得出一项更一般的结果：对于广义线性模型中基于非二次似然的方法，即 Barndorff-Nielsen 的 p*公式（Barndorff-Nielsen 1983年），其与局部均匀而非完全均匀先验下的贝叶斯后验分布渐近等价。这些分析结果确立了广义线性模型中非贝叶斯计算与贝叶斯计算的数值等价性，使我们能够在贝叶斯范式下开展进一步研究。

第 4 节我们提出一种效用函数，用于处理这样一个决策问题：在广义线性模型回归中，报告系数的方向性效应。该效用函数通过考量数据的统计分析师与委托相关分析的科学家客户之间的互动而得出。我们表明，这一方向性决策问题的结局不仅由π值决定，后者还衡量证据强度，因为它决定了该决策问题的完美信息期望值。

在第5节中，我们介绍一个用于科学发现可重复性的适当贝叶斯框架。从预测概率的角度出发，该框架使科学发现的使用者（例如政策制定者、监管机构，甚至公众）能够探究某项研究若被重复实施，其推断的稳健性。基于似然收敛为二次多项式这一假设，我们在本节推导了 p 的重复抽样分布（Goodman 1992年；Greenwald 等，1996年；Senn 2002；Boos 和 Stefanski 2011）以及 π-值的闭式表达。我们还说明了一种直接的蒙特卡洛方法，用于在对数似然的二次近似受到质疑时推导 p 和 π-值的重复抽样分布。对相关概率密度性质的数学考察，使我们得以批评使用重复试验所得的 p 和 π-值来衡量科学发现的可重复性。这种批评与其说基于哲学层面，不如说源于 p 和 π- 值变异性较大的技术问题。在第 6 节中，我们运用前几节建立的分析工具，考察钠 - 葡萄糖协同转运蛋白2抑制剂（SGLT2i）这一新兴心脏代谢药物类别的肾脏特异性结局的可重复性。此类药物具有广泛的肾脏获益（Johansen 和 Argyropoulos 2020年），但也有独特的副作用特征。我们利用这些试验阐释临床试验中的可重复性问题、先验分布的作用，以及广义线性模型回归近似的局限性。在最后一节中，我们尝试综合第2至第6节的内容，提出一种贝叶斯/似然混合方法，以

展示并解读考虑可重复性的科学研究结果。附录中演示了如何使用 R 语言编写相关计算程序。

2. 广义线性模型的似然方法

广义线性模型是分析连续型（线性回归）、二项型（逻辑回归）、离散型（多项回归）、发生率或事件发生时间（泊松回归）结局的广泛适用方法。这些技术广泛用于分析社会科学和医学领域的数据，而这两个领域正处于当前可重复性危机的前沿。广义线性模型的回归方法可提供效应度量，即对观察性研究和随机研究中的关联或治疗效应进行定量度量。本节旨在简要介绍基于似然的广义线性模型方法，为后续的贝叶斯分析奠定基础；对于希望深入了解广义线性模型的读者，可参阅以下教科书（McCullagh and Nelder 1989年；Lindsey 1997年；Pawitan 2013年；Wood 2017年；Lee 等 2017年；Dobson and Barnett 2018年）。

2.1. 广义线性模型的治疗效应估计

本质上，广义线性模型（GLM）实现了一个可分阶段求解的层级估计问题，具体如下：

- 阶段 1：建立可观测量与参数之间的关系，y，$\boldsymbol{\theta} = \theta_i, \quad i = 1..., \quad n$
- 阶段 2：规定参数之间关系的性质θ_i（抽样模型，若存在相关性则包含相关性）以及任何干扰参数（ϕ）
- 阶段 3：纳入有关关系一般形式的更多信息（链接函数—将在下文定义）、加性回归结构，以及通过投影到参数空间（$\boldsymbol{\beta}$）**进行的重新参数化**；该参数空间的维数小于可观测量与参数所在空间的维数 $\boldsymbol{\theta}$

在第 1 和第 2 阶段中，我们假设科学研究中收集的所有数据都来自属于指数族的某个分布，其密度表示如下：

$$f(y_i|\theta,\phi)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{\phi a_i^{-1}}+c(y_i,\phi)\right\} \quad \text{Eq. 1.}$$

参数 θ_i， ϕ 是位置参数和尺度参数，它们决定响应变量的均值和方差：$\mu_i = b'(\theta_i)$ 和 $\sigma_i^2 = b''(\theta_i)\phi a_i^{-1}$；而 GLM 的方差函数为$V(\mu_i) = V_i = b''(\theta_i)$。假设权重 a_i 已知，为简明起见，以下均取为 1，且不失一般性）。第 2 阶段通过指定一个抽样模型来完成，该模型通过观测值之间的相关性间接地将各个参数联系起来。对于独立同分布数据（i.i.d），各个观测值对问题的似然函数 （L）的贡献是相乘的。似然是问题中参数和数据的数学函数，数据为 $\boldsymbol{y}=\{y_1, y_2, \ldots, y_n\}$；对于 i.i.d 情形，其形式为：

$$L(\theta,\phi;\boldsymbol{y}|\mathbf{X})=\prod_{i=1}^{n}f(y_i|\theta,\phi)=\prod_{i=1}^{n}\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{\phi a_i^{-1}}+c(y_i,\phi)\right\}$$

广义线性模型在第 3 阶段通过在响应变量的均值中引入——连续可微变换而完成，该均值可假设服从关于协变量的线性模型：

$$\eta_i=g(\mu_i)=\boldsymbol{x}_i^T\boldsymbol{\beta}\,,\ i=1,2,\ldots,n\,,\ \dim(\boldsymbol{\beta})=p,\ \theta_i=(b')^{-1}(g^{-1}(\eta_i))$$

函数 $g(\cdot)$ 是广义线性模型的链接函数；当 $\theta_i=\eta_i$，即 $g(\cdot)=b^{-1}(\cdot)$ 时，该链接函数称为典范链接函数。

对于独立同分布的样本，将对数似然（*ll*）函数（似然的对数）视为 $\boldsymbol{\beta}$ 而非 θ 的函数，可表示为：

$$ll(\boldsymbol{\beta},\phi;\boldsymbol{y}|\mathbf{X})=\sum_{i=1}^{n}\left\{\frac{y_i\theta_i-b(\theta_i)}{\phi}+c(y_i,\phi)\right\} \quad \text{Eq. 2.}$$

通过求解得分方程，可以得到参数向量$\boldsymbol{\beta}$ 的估计值：

$$s(\beta_j) = \frac{\partial l}{\partial \beta_j} = 0 = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \times \frac{x_{ij}}{g'(\mu_i)}$$

$$W_i = \frac{w_i}{(\mu_i)\, V(\mu_i)\, g'(\mu_i)^2}, \quad w_i(\mu_i) = 1 + (y_i - \mu_i)\{V'(\mu_i)/V(\mu_i) + g'(\mu_i)/g'(\mu_i)\}, \quad G_i = g'(\mu_i)$$

$G_i = g'(\mu_i)$，对于典范连接函数，$V(\mu_i) = 1/g'(\mu_i)$，$V''(\mu_i) = -g''(\mu_i)/g'(\mu_i)^2$ 和 $w_i = 1$，$W_i = V(\mu_i)$，从而大大简化了得分方程的解析推导。从统计学视角来看，选择典范连接函数可使观测费舍尔信息矩阵（对数似然的海森矩阵）等于其在分布 $p(\boldsymbol{y}|\boldsymbol{\beta}, \phi)$ 下的期望值。采用迭代加权最小二乘算法求解非线性得分方程 $\mathbf{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{G}(\boldsymbol{y} - \boldsymbol{\mu})$，其中$\boldsymbol{W} = diag\{W_i\}$ 和 $\boldsymbol{G} = diag\{G_i\}$，以获得最大似然参数估计及其方差。请注意，在求解关于参数 $\boldsymbol{\beta}$ 的似然最大值时，尺度参数会从得分方程中约去。若将 IWLS 收敛时的估计值记为 $\widehat{\boldsymbol{\beta}}$，将回归模型的设计矩阵记为 $\mathbf{X}$ （通过按列堆叠行向量 $\boldsymbol{x}_i^T$ 构造），将 $\mathbf{W}^k = diag\{W_i^k\}$ 记为 IWLS 在第 k 次（收敛）迭代时的权重矩阵值，并将 $I(\widehat{\boldsymbol{\beta}})$ 记为观测费舍尔观测矩阵，则可得到最大似然估计的极限分布为多元正态分布（MVN）：

$$\widehat{\boldsymbol{\beta}}|\mathbf{X}, \boldsymbol{\beta}, \phi \to MVN\left(\boldsymbol{\beta}, I(\widehat{\boldsymbol{\beta}})^{-1}\right) = MVN(\boldsymbol{\beta}, \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} = \phi\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{W}^k\mathbf{X}\right)^{-1} \qquad \text{Eq. 3.}$$

为简化记法，我们将写作 $\mathbf{W}_{\widehat{\boldsymbol{\beta}}}$，而不是 $\mathbf{W}^k$。这种记法清楚地表明，权重矩阵以及观测信息矩阵都是 $\widehat{\boldsymbol{\beta}}$ 的函数。更一般地，可以定义矩阵函数，$\mathbf{W}(\boldsymbol{\beta}) = \mathbf{W}_{\boldsymbol{\beta}} = diag\left\{1 \;\; V(/\;\; \mu_i\; \boldsymbol{\beta} \;\; g'\; \mu_i\; \boldsymbol{\beta}\;\; {}^2 ()) (()) \right\}$，以表示任意点 $\boldsymbol{\beta}$ 处与对数似然相关的期望费希尔信息矩阵：

$$\mathcal{I}(\boldsymbol{\beta}) \equiv \left(E\left[-\frac{\partial^2 ll(\boldsymbol{\beta}, \phi|\boldsymbol{y})}{\partial \beta_j \partial \beta_l}\right]\right)_{j,k} = \phi^{-1}\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{W}_{\boldsymbol{\beta}}\mathbf{X}\right) \qquad \text{Eq. 4.}$$

在典范连接的情况下，$\mathbf{W}_{\boldsymbol{\beta}} = diag\{V(\mu_i(\boldsymbol{\beta}))\} = diag\{1/g'(\mu_i(\boldsymbol{\beta}))\} = diag\{b''(\theta_i)\}$而对于非典范链接，$\mathbf{W}_{\boldsymbol{\beta}} = diag\{b''(\theta_i)\}\, diag\{(d\theta_i/d\eta_i)^2\}$包含针对链接函数的调整。即使使用观测到的费希尔信息而非期望的费希尔信息，在适当重新定义权重矩阵的值后，协方差矩阵仍将保留公式4右侧给出的通用形式，从而形成标准差$\mathbf{W}_{\boldsymbol{\beta}}$（Wood 2017年）。

2.2. 尺度参数的估计

文献中提供了多种非贝叶斯计算方法，适用于离散参数（干扰参数）的值已知的情况，例如逻辑回归或泊松回归。尺度参数是干扰参数中最简单的例子，已有许多非贝叶斯方法用于计算估计量。我们将在下文介绍这些方法，然后在第3节中说明如何将这些计算方法推导为贝叶斯后验分布的（众数）近似。尺度参数的矩估计由以下方程给出：

$$\hat{\phi}_{MOM} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{V}_i} \qquad \text{Eq. 5.}$$

其中，$\hat{\mu}_i, \hat{V}_i = V(\hat{\mu}_i)$分别是广义线性模型中第 i 个观测的均值和方差函数，其参数取最大似然估计值。该估计量与伪似然 方法的估计量相同，后者假设皮尔逊残差 $r_{PL,i} = (y_i - \hat{\mu}_i)\sqrt{1/\hat{V}_i}$服从正态分布。

或者，也可以在扩展准似然（EQL）方法（Nelder 和 Pregibon，1987年）中使用广义线性模型的偏差来估计尺度参数。偏差是所考虑的约简模型与饱和模型的对数似然之差（经过缩放）；在饱和模型中，每个观测值 y_i 都有一个参数 θ_i 。若将 $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\boldsymbol{y}(\tilde{\boldsymbol{\beta}}))$ 和 $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\beta}}))$ 分别记为

饱和模型下规范参数的估计值和约简模型下规范参数的估计值，以及 $\widetilde{\boldsymbol{\beta}}$ 和 $\widehat{\boldsymbol{\beta}}$ 相应的回归参数，则偏差为：

$$\hat{\phi}_{EQL} = \frac{2\phi\{ll(\widetilde{\boldsymbol{\beta}}, \phi; \boldsymbol{y}|\mathbf{X}) - ll(\widehat{\boldsymbol{\beta}}, \phi; \boldsymbol{y}|\mathbf{X})\}}{n} = \frac{D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})}{n} = \phi \sum_{i=1}^{n} \frac{d(y_i, \widehat{\mu}_i)}{n} \quad \text{Eq. 6.}$$

$$d(y_i, \mu_i) = 2\{y_i(\widetilde{\theta}_i - \theta_i) - b(\widetilde{\theta}_i) + b(\theta_i)\} \quad \text{Eq. 7.}$$

量 $d(y_i,\ \widehat{\mu}_i)$等于缩放偏差残差的平方，可通过在最大似然估计值 $\widehat{\theta}_i$ 处计算式 7 得到。EQL 方法基于对广义线性模型族（式 1）的鞍点近似。该近似的出发点是将广义线性模型族改写为指数色散模型（Jorgensen，1987年，1997年）：

$$f(y_i) = h(y_i, \phi) \exp\left\{-\frac{1}{2\phi} d(y_i, \mu_i)\right\} \quad \text{Eq. 8.}$$

$$h(y_i, \phi) = \exp\big(c(y_i, \phi) + 2\phi\{y_i\widetilde{\theta}_i - b(\widetilde{\theta}_i)\}\big)$$

当 $\phi \to 0$ （小色散情形）时，在鞍点近似中，h(y_i, ϕ) $\approx \{2\pi\phi V(y_i)\}^{-1/2}$ （其中 $V(y_i)$ 是在 y_i 处计算的广义线性模型的方差函数）。鞍点近似可改写为广义线性模型的形式：

$$f(y_i) = \frac{\exp\left\{\dfrac{y_i\theta_i - b(\theta_i)}{\phi} - \dfrac{y_i\widetilde{\theta}_i - b(\widetilde{\theta}_i)}{\phi}\right\}}{\{2\pi\phi V(y_i)\}^{1/2}} \quad \text{Eq. 9.}$$

这种近似对于高斯分布和逆高斯分布是精确的。对于其他分布，这种近似的质量也出奇地好；例如，当泊松计数大于 3，或二项分布成功次数和失败次数大于 3 时，相对误差小于 3%（ Smyth 和 Verbyla，1999）。对于泊松和伽马广义线性模型，鞍点近似 consiste于将模型定义中出现的阶乘/伽马函数

概率质量函数替换为其斯特林公式近似。更一般地说，如果响应变量的均值距分布边界超过三个标准差，则这种近似大致成立。非正式地说，EQL近似将数据分解为两个组成部分："充分统计量"（$\widehat{\boldsymbol{\beta}}$）和"模型拟合优度的度量"（$D(\boldsymbol{y};\ \widehat{\boldsymbol{\mu}})$）。它还将模型的参数正交化，从而大大简化了模型拟合和参数估计所需的论述以及数值计算。

鞍点近似也可以计算尺度的偏差估计：$\hat{\phi}_{dev} \approx D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})\ (/\ n-p\ )$(Jorgensen 1987; Jørgensen 1992)。文献中提出的最后一种尺度参数估计基于修正剖面似然（Barndorff-Nielsen 1983; Jorgensen 1987）：

$$\hat{\phi}_{mPL} = \max_{\phi}\left(\frac{p}{2}\log(\phi) + \max_{\boldsymbol{\beta}|\phi} ll(\boldsymbol{\beta},\phi;\boldsymbol{y}|\mathbf{X})\right) \qquad \text{Eq. 10.}$$

这些不同估计量之间的关系此前已有讨论（Jorgensen 1987; Jørgensen 1987, 1992）：$\hat{\phi}_{mPL}$ 等价于$\hat{\phi}_{dev}$，类似地，在鞍点近似下，$\hat{\phi}_{MOM}$ 等价于$\hat{\phi}_{dev}$。有人提出，如果数值稳健性并非考虑因素，则优先采用$\hat{\phi}_{mPL}$ 而非矩估计量（Jørgensen 1992）。基于偏差的估计量在广义线性模型的 EQL 方法中发挥关键作用（Lee et al. 2017），但当离散参数趋于零或样本量趋于无穷时，其一致性尚不明确（Jorgensen 1987; Jørgensen 1987, 1992）。

2.3. 广义线性模型中的 p值与零假设显著性检验

非贝叶斯分析通常会在零假设显著性检验（NHST）中报告尾部概率（"p值"），以检验参数估计值是否等于特定值。p值是

获得至少与在零假设下观察到的结果同样极端的结果的概率 H_0。较小的p值表明，在零假设下观测到这些数据的可能性较低，因此可视为反对零假设的证据（Senn 2002；Berger 2003）。这类区域概率也受到阈值式决策分析的约束；小于统计显著性阈值（通常设为 $a = 00.05$）的p值会导致 拒绝 H_0。P – 值通过计算检验统计量得出。检验统计量是观测数据的函数，属于随机变量；随后，从已知的统计量累积分布函数中读取相应的概率。在零假设显著性检验中，通过将检验统计量与 H_0 对应于统计显著性阈值的临界统计量值进行比较，来作出关于 c_a 的决策。

在广义线性模型中，H_0 指出一组参数中的真实值（$\boldsymbol{\beta}$）等于某个预先指定的值。例如，可以指定$\boldsymbol{\beta}$中的所有元素（除单个元素β_i外）均等于 $\beta_{i,0}$。后者通常取为零，以表示治疗 效应对线性预测子以及广义线性模型的响应变量均无影响。计算广义线性模型检验统计量主要有两种方法：似然比检验（*LRT*），即备择假设下似然与零假设下似然之比的对数的两倍；以及 Wald 检验。如果对数似然恰好是二次函数，这两个统计量完全相同：在这种情况下，似然比检验基于零假设所隐含的参数值处的似然与最大似然（ML）处似然之间的“垂直”距离，而 Wald 检验基于两个参数值之间的距离（Buse 1982；Engle 1984）。对于由二次函数良好近似的规则对数似然，这两种检验渐近等价（这是广义线性模型似然方法的关键隐含假设之一）（Hayakawa 1975；Buse 1982；Cordeiro 1983, 2004；Engle 1984；Vargas等，2014）。鉴于两种检验之间的密切关系，下面将仅关注较为简单的（双侧）Wald 检验。单个参数的 Wald 检验p值和拒绝决策可写为：

$$p = 2\Phi(-|z_i|), reject\ H_0\ iff\ |z_i| > \Phi\left(1-\frac{a}{2}\right), z_i = \frac{\hat{\beta}_i - \beta_{i,0}}{se(\hat{\beta}_i)} \quad \text{Eq. 11.}$$

在此表达式中，Φ 是标准正态分布的累积密度，z_i 是 Wald 统计量平方根的值。当需要根据数据估计离散参数时，z_i 所服从的分布是学生t分布，而不是正态分布。$\hat{\beta}_i$的标准误$se(\hat{\beta}_i) = \sqrt{\boldsymbol{\Sigma}[i,i]}$，是协方差矩阵 $\boldsymbol{\Sigma} = \phi(\mathbf{X}^T\mathbf{W}_{\hat{\beta}}\mathbf{X})^{-1}$。 关于单变量 Wald 检验，有几点值得注意：首先，效应的方向性，即差值$\hat{\beta}_i - \beta_{i,0}$的符号，在p值的计算过程中实际上丢失了。此外，尽管检验本身针对的是一个精确的零假设，但其值可以根据两个互补方向事件的概率计算得出：

$$p = 2 \times min(\{P(\hat{\beta}_i \geq \beta_i | \beta_i = \beta_{i,0}), P(\hat{\beta}_i < \beta_i | \beta_i = \beta_{i,0})\}) \quad \text{Eq. 12.}$$

最后，如果最不可能的方向性事件的概率小于 $a/2$，则拒绝 H_0。等价地，如果最可能的方向性事件的概率大于$1-a/2$，则会拒绝 H_0。从这个非常特定的意义上说，可以将 p值 解读为提供了支持方向性协变量效应的证据，因为拒绝零假设最终是由方向效应的概率决定的。尽管对 p 值计算的分析看起来很复杂，但它揭示了报告这项检验时“幕后”发生的情况。

3. 广义线性模型的贝叶斯方法

在广义线性模型的贝叶斯分析中，必须为未知的回归参数和离散参数（如果未知）指定先验分布。在本节中，我们将论证，广义线性模型贝叶斯分析的默认先验分布应当是一种可能为非正常分布的分布：它在参数向量 $\boldsymbol{\beta}$ 所在的超立方体区域内为常数，并且与尺度参数 ϕ 成反比。因此，$\boldsymbol{\beta}$ 和尺度参数的先验分布相互独立；这种选择是合理的，因为这些参数分别决定响应变量尺度的位置（Box 和 Tiao 1973）。尽管这一提议不同于联合非信息先验（Jeffreys 1946），但它务实地承认，$\boldsymbol{\beta}, \phi$ 的杰弗里斯先验"在解析计算中相当难以处理，数值计算中可能也不实用"（Ibrahim 和 Laud 1991），而因子分解为 $p(\boldsymbol{\beta}, \phi) = p(\boldsymbol{\beta}) \times p(\phi)$ 的先验分布则可能相当容易处理。若分别为决定位置的参数（$\boldsymbol{\beta}$）和影响尺度的参数（ϕ）选取信息性先验，我们便得到了杰弗里斯为客观最小信息先验提出的修正方案（Jeffreys 1946；Box 和 Tiao 1973）。在采用 $p(\boldsymbol{\beta})$ 这一设定时，可以得到贝叶斯分析与频率学派分析在数值上的等价；不同的方差参数估计量则源于对 ϕ 的先验分布作出的不同选择。因此，我们将在以下各节中详细讨论这一参考先验。

3.1. 治疗效应的有界均匀先验

我们对 $\boldsymbol{\beta}$ 先验分布的具体选择，可从三个不同角度加以论证：有限世界论证；将其视为一种为函数关系诱导出均匀先验分布的方法（Bornkamp 2012）；以及将其视为针对研究者群体对主导似然／局部均匀先验方法（Box 和 Tiao 1973）的适当扩展。

3.1.1. 有限世界论证

广义线性模型常用于分析现实世界现象的属性。这样的使用背景立即显现出统计模型的空间（可能是无限的）与通过对物理世界进行抽样而获得的测量值具有有限性之间的矛盾。由于广义线性模型参数可与实际测量值直接对应，因此可以将每个广义线性模型参数的允许取值范围限制在一个有限区间内，从而化解这一表面上的矛盾。该区间的长度可通过考察测量值／现实某一方面的有限范围来推导，而该方面在广义线性模型中被建模为响应变量。通过考察这些量的范围，可以推导出参数范围的有用界限，进而为选择合适的"非信息性"先验分布提供参考。为说明这一过程，考虑一个取值被限制在区间 $\left[y_1^{min}, y_1^{max}\right]$ 内的单一结果变量 y_1 ，以及一个对应于广义线性模型截距项的单一参数 β_1。那么：

$$y_1^{min} \le y_1 \le y_1^{max} \Rightarrow \beta_1^{min} = g^{-1}(y_1^{min}) \le \beta_1 \le g^{-1}(y_1^{max}) = \beta_1^{max}$$

通过考虑广义线性模型中设计矩阵内各变量的有限取值范围 $\left[X_i^{min}, X_i^{max}\right]$，并求解以下不等式，可以类似地推导出广义线性模型其他参数取值范围的限制：

$$\beta_i^{min} = \frac{g^{-1}(y_1^{min})}{X_1^{min}} \le \beta_i \le \frac{g^{-1}(y_1^{max})}{X_1^{max}}$$

在限定了每个参数的取值范围（假设它是模型设计矩阵中唯一出现的参数）后，接下来引入均匀先验：

$$p(\beta_1, \beta_2, \dots, \beta_p) = p^{-1} \prod_{i=1}^{p} p(\beta_i) = p^{-1} \prod_{i=1}^{p} \frac{[\![\beta_i \in \left[\beta_i^{min}, \beta_i^{max}\right]]\!]}{\beta_i^{max} - \beta_i^{min}}$$

按设计矩阵中参数的数量进行缩放，可确保广义线性模型的线性预测子（它是所有参数的函数）始终被限制在区间 $\left[g^{-1}\left(y_1^{min}\right), g^{-1}(y_1^{max})\right]$ 内。这种缩放还能避免内部不一致，例如违反区间 $\left[y_1^{min}, y_1^{max}\right]$ 的界限（如果只逐一考察变量）。特别是，无法保证 $\sum_i X_i^{min}\beta_i^{min} \geq y^{min} \wedge \sum_i X_i^{max}\beta_i^{max} \leq y^{max}$，实际上，除非先验分布按参数数量进行缩放，否则这两个陈述几乎总会同时不成立。艾弗森括号表示法 $[\![\beta_i \in \left[\beta_i^{min}, \beta_i^{max}\right]]\!]$表示，如果任一参数超出其界限，先验分布的取值就为零，并将先验分布限制为超立方体上的均匀分布。尽管将先验分布限制在超立方体内看似降低了先验分布的通用性，但应注意，这种限制隐含于任何数值计算中，因为数值计算的精度有限。事实上，只需援引这一论点，就能将先验分布的范围限定为所用算术库和浮点精度所能表示的数值范围。

3.1.2. 广义线性模型的函数型均匀先验

我们也可以从不同的角度得出相同的均匀先验（即在超立方体上均匀），也就是寻求在通过广义线性模型回归估计的函数形状空间上均匀的先验分布（Ghosal 等人，1997年；Bornkamp，2012年）。这类先验分布通过可微映射 φ: $\boldsymbol{B} \longrightarrow M$，将参数$\boldsymbol{\beta}$ 从紧致子空间 $\boldsymbol{B} \in$ 实数集^p 映射到另一个配备度量 d的紧致度量空间 (M, d)。假设度量 $d(\varphi, \varphi_0) = d\left(\varphi(\boldsymbol{\beta}), \varphi(\boldsymbol{\beta_0})\right) = d^*(\boldsymbol{\beta}, \boldsymbol{\beta_0})$具有足够良好的性质，可以在局部用马氏距离近似，即：

$$d^*(\boldsymbol{\beta}, \boldsymbol{\beta_0}) = c_1\sqrt{c_2(\boldsymbol{\beta} - \boldsymbol{\beta_0})^T \boldsymbol{U}(\boldsymbol{\beta_0})(\boldsymbol{\beta} - \boldsymbol{\beta_0}) + O(\|\boldsymbol{\beta} - \boldsymbol{\beta_0}\|^k)} \qquad \text{Eq. 13.}$$

其中 $\boldsymbol{U}(\boldsymbol{\beta_0})$ 是一个特征值严格为正的对称矩阵，$k \geq 3$ 和 c_1， c_2 是正常数。在 $(M,\ d)$ 上施加均匀分布，会在 $\boldsymbol{B}$ **上诱导出一个可能非均匀的分布：**

$$p(\boldsymbol{\beta}) = \frac{\sqrt{|\boldsymbol{U}(\boldsymbol{\beta})|}}{\int_{\boldsymbol{B}} \sqrt{|\boldsymbol{U}(\boldsymbol{\beta})|} d\boldsymbol{\beta}}, \qquad |\boldsymbol{U}(\boldsymbol{\beta})| \equiv \det \boldsymbol{U}(\boldsymbol{\beta}) \tag{Eq. 14.}$$

对于广义线性模型，我们可以要求单个观测的线性预测子尺度服从均匀分布，也就是说，在欧几里得距离度量 $d^*(\boldsymbol{\beta},\ \boldsymbol{\beta_0}) = \sqrt{(\boldsymbol{x^T\beta} - \boldsymbol{x^T\beta_0})^T(\boldsymbol{x^T\beta} - \boldsymbol{x^T\beta_0})}$下，映射 φ 是 $\eta = \boldsymbol{x^T\beta}$ 。这些选择会导出一个常数 $\boldsymbol{U}(\boldsymbol{\beta_0}) = \boldsymbol{xx^T}$，因此在具有任意但有限界限的超立方体上得到常数先验分布。注意，~~先验分布 $\propto \sqrt{\det \boldsymbol{xx^T}}$ 虽然与~~广义线性模型的杰弗里斯先验密切相关（Ibrahim 和 Laud 1991），但只有在高斯情形下才与之重合。事实上，杰弗里斯先验本身可视为统计模型残差密度空间上的函数型均匀先验，所依据的度量为海林格距离或库尔贝克-莱布勒距离（Balasubramanian 1997；Ghosal 等，1997；Bornkamp 2012）。通过对一系列不断扩大的紧空间取极限，可以将其扩展到整个 R^p （Ghosal 等，1997）；然而，由于无法保证分母中的积分收敛，所得先验密度可能不可积。

3.1.3. 研究者群体的似然支配先验分布。

似然占主导的先验分布，是指数据的似然比先验分布更集中（“尖峰”），而先验分布在似然具有显著值的区域内几乎不变（“平坦”），并且在该区域之外不会取很大的值。这种情形对应于科学推断中的“稳定估计”原则（Edwards 等（1963））。采用这类先验分布有两项主要依据（Box 和 Tiao，1973年）。第一，除非科学研究有望增进相对于研究开展之前已有认识的知识，否则通常不会开展。因此，在这类情形中采用的先验分布对最终推断的影响应尽可能小。第二，科学论述从来不是孤立进行的：科学家会利用实验数据说服同事接受关于世界状态的看法。后者的观点往往与开展研究的科学家不同。这种意见分歧可能如此之大，以至于一个

直接向研究人员征集的信息性先验分布会在观测数据的潜在取值范围内广泛分散。在这两种情形下，应如何进行贝叶斯推断？一个直接的答案是选择默认的参考先验分布，它代表了这样一种观点：某人事先对世界状态知之甚少。此人的后验分布将由似然（“数据”）主导，而非先验分布，因此其不受任何人在数据可用之前可能持有的想法影响。此类先验分布确有必要，因为在某些情形下（例如药物或器械的监管审批），“往往需要呈现出客观性的表象”（Berger 2006）。在监管背景之外，科学家常借助统计学为其研究发现提供客观验证；在这类情形下，客观贝叶斯先验分布能够实现“易于理解地传达观测数据中的信息，这些信息是通过统计模型传达的”，同时确保特定问题的答案以手头的数据为条件，并“尊重频率学派观点，即方法必须确保科学家在重复使用时能够取得成功”（Berger 2006）。

在似然函数具有显著值的范围内，局部均匀先验，即恒定或近乎恒定的先验（Edwards 等（1963）），可作为默认先验的基础。此外，如果该先验不会强烈偏向似然函数的其他区域，那么它就符合由似然主导的先验的标准。只要将边界设得足够宽，不会“截断”似然函数具有显著值的区域，超立方体上的均匀先验就同时符合这两项标准。尽管这些可追溯至 1960 年代的考量表明，均匀先验可作为研究者群体的参考选择，但导出均匀先验的数学论证并不直观。通过考察科学家意见多样性的正式数学表示（“先验信息征询”），可以使这一论证更加清晰。

3.1.3.1. 先验信息征询与局部均匀先验。自线性回归模型的早期研究（Kadane 等（1980））以来，已有大量出版物探讨广义线性模型的先验信息征询（Bedrick 等（1996）；Denham 和 Mengersen（2007）；Garthwaite 等（2013）；Hosack 等（2017））。在这些预测方法中，研究人员会询问专家对潜在可观测量（例如响应变量的条件均值）在不同设计点上的看法。随后，研究人员采用合适的回归模型分析专家意见，例如（加权或非加权）最小二乘法（Kadane 等（1980）；Denham 和 Mengersen（2007）；Hosack 等（2017）），甚至广义线性模型回归（Bedrick 等（1996））。这种方法避开了结构方法所要求的难度大得多的认知任务；在结构方法中，需要征询专家对回归参数 β_i 本身取值的看法（Kadane 和 Wolfson（1998））。从专家意见中导出先验分布的不同程序，最终往往会将其表示为多元高斯分布（Racine 等（1986）；Grieve（1988）；Bedrick 等（1996）；Denham 和 Mengersen（2007）；Garthwaite 等（2013）；Hosack 等（2017））。因此，如果要向群体征询先验分布，该先验就会呈现如下形式：

$\boldsymbol{\beta}|\boldsymbol{\beta}_{prior}, \boldsymbol{\Sigma}_{\text{prior}} \sim MVN\left(\boldsymbol{\beta}_{prior}, \boldsymbol{\Sigma}_{prior}\right), \boldsymbol{\Sigma}_{\text{prior}} = \boldsymbol{\sigma}_{prior}\boldsymbol{C}_{prior}\boldsymbol{\sigma}_{prior}$在后一个方程中，超参数 $\boldsymbol{\beta}_{prior}, \boldsymbol{\Sigma}_{prior}$ 将固定为通过信息征询获得的值。该先验分布的协方差也可以通过边际方差 $\boldsymbol{\sigma}_{prior} = \sqrt{diag\left(\boldsymbol{\Sigma}_{prior}\right)}$和相关矩阵 $\boldsymbol{C}_{prior}$进行等价参数化。

若没有开展正式 elicitation 活动，我们只能获得群体先验分布的数学形式。现在，问题已转化为指定多元高斯分布的众数、边际方差和相关矩阵，使其在潜在数据/似然的取值范围内尽可能弥散。我们使用公式 15 逐步构建这种“局部平坦”的先验分布：首先，我们要探讨能否从数学上定义局部均匀性的概念。我们认为，可以肯定地回答这个问题，即通过以下方式定义这一概念：

Radon-Nikodym导数，相对于一个 σ-有限测度，例如欧几里得空间 R^p上的 Lebesgue测度。如果公式 15 所隐含的概率测度相对于 Lebesgue测度的导数在 R^p上为常数（概率为零的集合除外）（Campbell 1966），那么它在所有可测集上都是局部均匀的（包括 $\boldsymbol{B}$），这些集合的并集构成 R^p。通过 Radon-Nikodym 导数定义局部均匀性，还有一个不太显而易见的好处：这些导数也可用作与公式 15 相对应的概率测度取值范围的广义测度。通俗地说，一种"延伸"得比另一种更远的测度，将覆盖参数空间 $\boldsymbol{B}$中的大片区域；由于相应的密度核可积，将其分散到广阔区域，就能确保先验分布在整个空间中都取较小的数值。

为了确定局部平坦高斯分布的协方差矩阵，我们注意到，归一化常数或微分熵（Campbell 1966；Peña and Rodrı´guez 2003；Chen et al. 2016）都是衡量分布"离散程度/范围/广度"的现成标量指标标量：前者是密度核的黎曼积分，而后者是概率测度相对于与该概率测度对应的密度的 Radon-Nikodym 导数的负对数的 Lebesgue 积分。归一化常数和微分熵都是多元正态分布的协方差矩阵行列式（即"广义方差"（Wilks 1932））的增函数。将 Hadamard不等式应用于行列式可得 $\left|\boldsymbol{\Sigma}_{\text{prior}}\right| \ni \left|\boldsymbol{C}_{prior}\right|\left|\boldsymbol{\sigma}_{prior}\right|^{\mathbf{2}} \leq \left|\boldsymbol{\sigma}_{prior}\right|^{\mathbf{2}}$，即一般多元正态分布的离散程度将小于具有相同边际方差但各分量相互独立的另一个多元高斯分布。因此，通过多元高斯分布定义局部均匀先验分布的一个重要首要步骤，是选择单位矩阵作为相关矩阵，从而使分布尽可能广泛地扩散。此外，高斯分布只有一个众数，因此，如果将先验分布的众数设定在参数空间内部（似然峰值很可能出现的位置），那么先验分布不仅会受到似然的支配，而且随着参数向……移动，也不会取较大的值。

边界$\boldsymbol{B}$。因此，将高斯分布的众数设在$\boldsymbol{B}$ 内部的任意位置，都有助于确保局部均匀性。

3.1.3.2. 检验先验和探索先验。

根据前两段中的启发式论证，我们现在可以集中讨论如何在 实数集 的子集上局部定义先验分布，并将参数向量 $\boldsymbol{\beta}$ 的每个高斯分量 β_i 与其他所有分量独立地加以考虑。此时，我们将区分两种可能性：检验型先验，其中高斯先验分布 $\beta_{i,0}^{prior}$ 的众数具有特殊意义，需要明确指定：

$$p_{test}\left(\beta_i|\beta_{i,0}^{prior},\sigma_{i,0}^{prior}\right)\sim N\left(\beta_i|\beta_{i,0}^{prior},\sigma_{i,0}^{prior}\right)$$

以及探索型先验，其中众数本身通过在包含该区间的闭区间 $\left[\beta_i^{min},\beta_i^{max}\right]$ 上为其指定超先验而进行概率化处理，即“更大”，β_i 很可能位于该区间内：

$$p_{explore}\left(\beta_i|\beta_{i,0}^{prior},\sigma_{i,0}^{prior}\right)=\int_{\beta_i^{min}}^{\beta_i^{max}}N\left(\beta_i|\beta_{i,0}^{prior},\sigma_{i,0}^{prior}\right)p\left(\beta_{i,0}^{prior}\right)d\beta_{i,0}^{prior}$$

相关区间本身可以通过“有限-世界”论证来指定，如第3.1.1节所述，而对$\beta_{i,0}^{prior}$ 而言，最简单的超先验分布是在区间$\left[\beta_{i,0}^{min},\beta_{i,0}^{max}\right]$上的均匀分布。对于检验型先验和探索先验，先验分布$\sigma_{i,0}^{prior}$的标准差$\beta_i$ 也应以符合局部均匀性的方式加以指定。最直接的选择是将$\sigma_{i,0}^{prior}$ 设为一个固定的大数，其中“较大”的含义可操作化为不影响数值计算的最大值。或者，也可以将标准差视为干扰参数，并使用适当的超先验分布对其进行积分消去。采用与广义线性模型（第3.2节）中尺度参数的先验分布类似的论证，我们可以设定超先验分布$p(\sigma_{i,0}^{prior})=$

$[\![\sigma_{i,\ 0}^{\ prior} \in [\sigma_{i,0}^{min}, \sigma_{i,0}^{max}]]\!]$ $(\sigma_{i,\ 0}^{\ max} - \sigma_{i,\ 0}^{\ min})$ /，从而得到一个非常规的检验型先验分布，其对于β_i 的先验概率与两个不完全伽马函数[1]之差成正比：

$$\left[\Gamma\left(0, \frac{\left(\beta_i - \beta_{i,0}^{prior}\right)^2}{2\sigma_{i,0}^{max\,2}}\right) - \Gamma\left(0, \frac{\left(\beta_i - \beta_{i,0}^{prior}\right)^2}{2\,\sigma_{i,0}^{min\,2}}\right)\right] \Big/ \left[2\sqrt{2\,\pi}\left(\sigma_{i,0}^{max} - \sigma_{i,0}^{min}\right)\right]$$

方差参数的超先验更常见的选择是缩放逆卡方分布，由此得到针对 β_i 、尺度为 s_i 且具有 ν_0 个自由度的 Student-t 检验型先验。相应的探索型先验可以通过从检验型先验中积分消去 $\beta_{i,\ 0}^{\ prior}$ 而导出。六种检验型和探索型先验的未归一化形式见表 1。它们都是单峰的，并且在 $\beta_i = \beta_{i,\ 0}^{\ prior}$ 时取得最大值；探索型先验仅在子集 $[\beta_{i,0}^{min}, \beta_{i,0}^{max}]$ 中非零。在点 $\beta_{i,\ 0}^{\ min} < \beta_i < \beta_{i,\ 0}^{\ max}$ 处的值与最大值之间的相对差异，可以通过适当选择超先验而任意缩小。

表 *1* 根据方差参数的超先验划分的检验型或探索型先验的未归一化密度

超先验 方差的 $\sigma_{i,0}^{prior}$	**检验型先验**	**探索型先验†**
固定值	$N(\beta_i \mid \beta_{i,0}^{prior}, \sigma_{i,0}^{prior})$	$0.5 \times (\Phi(\frac{\beta_i - \beta_{i\,0}^{\ max}}{\sqrt{2}\sigma_{i\,0}^{\ prior}}) - \Phi(\frac{\beta_i - \beta_{i\,0}^{\ min}}{\sqrt{2}\sigma_{i\,0}^{\ prior}}))$
均匀分布于 $[\sigma_{i,0}^{min},\ \sigma_{i,0}^{max}]$	$\Gamma\left(0, \frac{(\beta_i - \beta_{i,0}^{prior})_2}{2\sigma_{i,0}^{max\,2}}\right) - \Gamma\left(0, \frac{(\beta_i - \beta_{i,0}^{prior})_2}{2\,\sigma_{i,0}^{min\,2}}\right)$	$\int_{\beta_{i,0}^{min}}^{\beta_i} \left[\Gamma\left(0, \frac{(\beta_i - \beta_{i,0}^{prior})2}{2\sigma_{i,0}^{max\,2}}\right) - \Gamma\left(0, \frac{(\beta_i - \beta_{i,0}^{prior})2}{2\,\sigma_{i,0}^{min\,2}}\right)\right] d\beta_i$ ǂ
Inv-$\chi^2(\nu_0,\ s_i^2)$	$St(\beta_i \mid \beta_{i,0}^{prior}, \sigma_{i,0}^{prior}, \nu_0)$	$\int_{\beta_{i,0}^{min}}^{\beta_{i,0}^{max}} St(\beta_i \mid \beta_{i,\ 0}^{\ prior}, \sigma_{i,\ 0}^{\ prior}, \nu_0) d\beta_i$ ¶

*†*这些先验分布恒为零 $[\![\beta_i^0 \notin [\beta_{i,0}^{min}, \beta_{i,0}^{max}]]\!]$，ǂ该积分可利用误差函数和不完全伽马函数的代数表达式以闭式计算，¶该积分可利用高斯超几何函数 *2F1* 以闭式计算

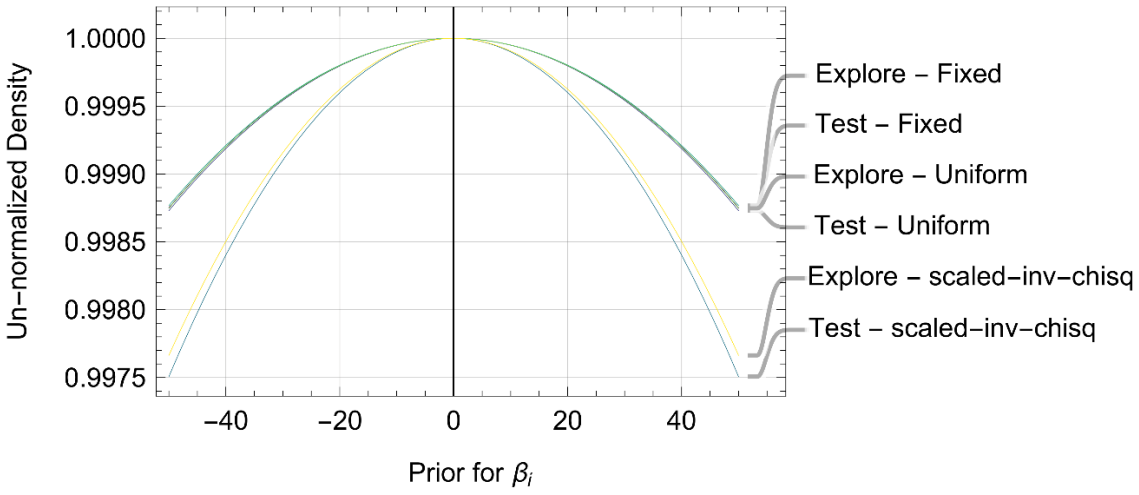
我们以六种先验分布中单个参数的情况为例，在图 *1* 中以图形方式展示这些要点

表 1。为了生成这幅图，我们将 $\sigma_{i,\ 0}^{\ prior} = 1000$ （一个常见的“默认”

[1] 该公式假设 $\beta_i \neq \beta_{i,0}^{prior}$，否则不完全伽马函数无定义。对于 $\beta_i = \beta_{i,0}^{prior}$，其密度值可用闭式表示为 $\log(\sigma_{i,\ 0}^{\ max}/\sigma_{i,0}^{min})/[\sqrt{2\,\pi}(\sigma_{i,0}^{max} - \sigma_{i,0}^{min})]$。

固定值先验的一种非信息性选项），$\beta_{i,0}^{min} = -200$，$\beta_{i,0}^{max} = +200$，$\beta_{i,0}^{prior} = 0$，$\sigma_{i,0}^{min}$=900，$\sigma_{i,0}^{max} = 1100$，并将 Student t 分布的自由度设为 1。所得先验分布在区间 $[-50,50]$ 内的变化很小（不到 0.25%），因此在该区间内近似均匀。尽管这些先验分布的函数形式迥异，但相对于达到的最大值而言，它们在局部的行为相当相似，也就是说，它们都类似于一条“平坦”的抛物线。事实上，可以通过选择超参数，使这些先验分布的形状不仅在 $[\beta_{i,0}^{min}, \beta_{i,0}^{max}]$ 中变得完全相同，而且在同一区间内，先验分布本身也可以视为恒定。

图 1 *x2* 单个参数的检验型先验分布与探索型先验分布的比较。



3.2. 广义线性模型尺度参数的先验分布

GLM 的尺度参数先验分布可以通过注意到该参数对于 GLM 的响应变量而言是一个真正的尺度参数来推导。它还用于缩放似然中的不确定性，这一点在考察公式 9 中的鞍点近似和公式 3 中给出的最大似然解时尤为明显。在其他情境下，例如线性回归或位置-尺度模型中，尺度参数的非信息先验被视为对数中的常数，或与尺度参数的未变换值成反比：$p(\log(\phi)) \propto c \Rightarrow p(\phi) \propto \phi^{-1}$ 。这也称为

尺度参数的杰弗里斯先验。与 $\boldsymbol{\beta}$ 的情形一样，我们可以将尺度参数的取值范围限制为实数集的真子集：

$$0 \leq \phi_{min} \leq \phi \leq \phi_{max} < \infty$$

在本研究中，我们将取 $\phi_{min} = 0$，因此均匀先验将等于 $\phi = \phi_{max}^{-1}$。或者，我们也可以沿用 $\boldsymbol{\beta}$ 的思路，将 ϕ 的先验限制为在区间 $[\phi_{min}, \phi_{max}]$，$p(\phi) \propto c$ 上服从均匀分布。我们之所以作出这一选择，是因为在有限世界中，没有任何事物——甚至连变化本身——是无限的。

3.3. 贝叶斯分析与似然主义分析之间的数值等价

3.3.1. 已知广义线性模型尺度参数的情形

在超立方体上采用均匀先验时，将贝叶斯定理应用于广义线性模型回归建模十分直接：

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}, \phi) = \frac{p^{-1} \exp\left(\sum_{i=1}^{n} \frac{y_i \boldsymbol{x}_i' \boldsymbol{\beta} - b(\boldsymbol{x}_i' \boldsymbol{\beta})}{\alpha_i(\phi)} + c(y_i, \phi)\right) \prod_{i=1}^{p} \frac{[\![\beta_i \in [\beta_i^{min}, \beta_i^{max}]]\!]}{\beta_i^{max} - \beta_i^{min}}}{p^{-1} \int \exp\left(\sum_{i=1}^{n} \frac{\boldsymbol{x}_i' \boldsymbol{\beta} - b(\boldsymbol{x}_i' \boldsymbol{\beta})}{\alpha_i(\phi)} + c(y_i, \phi)\right) \prod_{i=1}^{p} \frac{[\![b_i \in [\beta_i^{min}, \beta_i^{max}]]\!]}{\beta_i^{max} - \beta_i^{min}} d\boldsymbol{\beta}}$$

$$= \frac{\exp(ll(\boldsymbol{\beta}, \phi; \boldsymbol{y}|\mathbf{X}))\, [\![\beta_i \in [\beta_i^{min}, \beta_i^{max}]]\!]}{\int \exp(ll(\boldsymbol{\beta}, \phi; \boldsymbol{y}|\mathbf{X}))[\![b_i \in [\beta_i^{min}, \beta_i^{max}]]\!]\, d\boldsymbol{\beta}} \approx \frac{\exp(ll(\boldsymbol{\beta}, \phi; \boldsymbol{y}|\mathbf{X}))}{\int \exp(ll(\boldsymbol{\beta}, \phi; \boldsymbol{y}|\mathbf{X}))\, d\boldsymbol{\beta}}$$

最后一行的近似等式成立的条件是：超立方体的边界延伸到似然具有显著值的区域之外，这样在通过拉普拉斯近似计算这些积分时，将积分上下限延伸至无穷大所引入的误差会呈指数级地小。关于这一计算，有几点值得注意：a) 只要参数的先验分布有界且在参数空间内为常数，先验分布就无关紧要；b) 参数的界限只有在限制积分范围时才会产生影响，

分母中的归一化常数，并将每个参数限制在用于定义先验分布的超立方体内。

为了将贝叶斯 $\exp(ll(\boldsymbol{\beta},\phi;\boldsymbol{y}|\mathbf{X}))/\int \quad ll\ \boldsymbol{\beta}\ \phi\ \boldsymbol{y}|\mathbf{X}\quad d\boldsymbol{\beta}\exp((,;))$ 与似然主义分析联系起来，我们必须考察后者赖以运作的机制。随后，我们可以把注意力转向贝叶斯解中的似然（分子）和分母中的多元积分，并推导出与似然主义分析假设相容的表达式。这样便能确保两种解的推导基于相同的数学假设。在计算上，似然主义分析只使用对数似然的最大值所在位置及该点附近的曲率。只使用这些信息的近似贝叶斯解可以通过以下方式得到：a）在似然最大值处对分子作多元泰勒展开，并仅保留前两个非零项（常数项和二次项）；b）通过分母中的多元拉普拉斯近似计算归一化常数。与最大似然方法类似，拉普拉斯近似利用被积函数的最大值及最大值附近的曲率，来近似指数积分的值（Fulks 1960年；Fulks and Sather 1961）。只要最大值位于参数空间边界以内，我们就可以写成：

$$\begin{aligned}
\exp(ll(\boldsymbol{\beta},\phi;\boldsymbol{y}|\mathbf{X})) &\approx \exp\left(ll(\widehat{\boldsymbol{\beta}},\phi|\boldsymbol{y})-\frac{1}{2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right) \\
\int \exp(ll(\boldsymbol{\beta},\phi;\boldsymbol{y}|\mathbf{X}))\,d\boldsymbol{\beta} &\approx (2\pi)^{\frac{p}{2}}\sqrt{|\boldsymbol{\Sigma}|}\exp\left(ll(\widehat{\boldsymbol{\beta}},\phi;\boldsymbol{y}|\mathbf{X})\right) \\
p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi) &\approx MVN(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}},\boldsymbol{\Sigma})
\end{aligned} \qquad \text{Eq. 16.}$$

在第二行和第三行中，我们利用了广义线性模型采用典范连接时，期望费希尔信息矩阵与观测费希尔信息矩阵在数值上的相等性。对于非典范链接，随着样本量增加，观测费希尔信息与期望费希尔信息仅在渐近意义上等价。因此，这一推导确立了贝叶斯分析与似然主义近似分析在**数值上**的等价性。在这两种情况下，正态分布的曲率（它决定了

概率在峰值附近的集中程度）完全相同，且可根据观测费希尔信息计算得出。需要指出的是，通过考察广义线性模型对数似然的性质，可以使这一高度启发式的推导严谨得多（Kass 等人（1990））。相关假设包括：被积函数具有六阶连续可微性、设计矩阵满秩，且方差参数远离边界。

3.3.2. 未知位置参数和尺度参数的情形

此情形下的后验分布可写为：

$$p(\boldsymbol{\beta}, \phi|\boldsymbol{y}, \mathbf{X}) = \frac{\exp(ll(\boldsymbol{\beta}, \phi; \boldsymbol{y}|\mathbf{X}))\, p(\phi)}{\iint \exp(ll(\boldsymbol{\beta}, \phi; \boldsymbol{y}|\mathbf{X}))p(\phi)\, d\boldsymbol{\beta}\, d\phi} \qquad \text{Eq. 17.}$$

关于 $\boldsymbol{\beta}$ 的推断（通常是感兴趣的参数）可通过积分消去干扰参数 ϕ 来进行：

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}) = \int p(\boldsymbol{\beta}, \phi|\boldsymbol{y}, \mathbf{X}) d\phi = \int p(\boldsymbol{\beta}|\phi, \boldsymbol{y}, \mathbf{X}) p(\phi|\boldsymbol{y}, \mathbf{X}) d\phi \qquad \text{Eq. 18.}$$

其中 $p(\phi|\boldsymbol{y}, \mathbf{X}) = \int p(\boldsymbol{\beta}, \phi|\boldsymbol{y}, \mathbf{X})\, d\boldsymbol{\beta}$。在详细讨论贝叶斯估计之前，我们将首先给出一种启发式推导，以概括我们计算的核心思想。特别值得注意的是，在广义线性模型中，ϕ 的维度远少于$\boldsymbol{\beta}$，也就是说，它是一个标量。给定大量数据y， $p(\phi|\boldsymbol{y}, \mathbf{X})$通常会呈现相对尖锐的峰值，这意味着$p(\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}) \approx p(\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}, \hat{\phi})$，其中$\hat{\phi}$ 是ϕ 的边际众数——只要先验分布$p(\phi)$相当无信息或“平坦”，它或多或少就等同于基于$p(\boldsymbol{y}|\mathbf{X}, \phi)$得到的$\phi$ 的最大似然估计。

该近似 $p(\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}, \hat{\phi})$ 具有如下形式：

1. 如果 ϕ 是通过分层方式估计的：$p(\boldsymbol{y}|\mathbf{X}, \phi) = \int p(y|\mathbf{X}, \boldsymbol{\beta}, \phi)\, p(\boldsymbol{\beta}|\phi)\, d\boldsymbol{\beta}$

   并使用 贝叶斯定理中先验分布的平坦程度以及众数附近的尖锐程度 ：

$$p(\phi|\boldsymbol{y}, \mathbf{X}) \propto p(\boldsymbol{y}|\mathbf{X}, \phi)p(\phi) = p(\boldsymbol{y}|\mathbf{X}, \phi) \approx p\left(\boldsymbol{y}|\mathbf{X}, \hat{\phi}\right)$$

2. 的后验分布 $\boldsymbol{\beta}$ 根据先验分布 $p(\boldsymbol{\beta}|\hat{\phi})$，按照常规方法计算得出

这就好像是"先验分布是根据数据估计出来的"——这通常被称为"经验贝叶斯"，是对严格贝叶斯分析的一种近似。

贝叶斯近似与频率学派近似之间的数值等价，可以通过对某种算法进行变形而更正式地得到：该算法通过以下方式之一计算贝叶斯积分：a）对于峰值尖锐的被积函数，使用狄拉克δ函数进行卷积；或 b）对于峰值不那么尖锐、且可用多元二次多项式近似的指数化对数似然，使用拉普拉斯方法：

1. 首先，对于任意给定的 ϕ，我们围绕 $\widehat{\boldsymbol{\beta}}$ 的最大似然估计值展开对数似然，在分子和分母中均如此，从而将后验密度分解为两部分：$\widehat{\boldsymbol{\beta}}$ 处的对数似然值作为 ϕ 的函数，以及依赖于 $\boldsymbol{\beta}$ 和 ϕ 的多元正态密度核。符号表示如下：

$$p(\boldsymbol{\beta},\phi|\boldsymbol{y},\mathbf{X}) \approx \frac{\exp\left(ll(\widehat{\boldsymbol{\beta}},\phi;\boldsymbol{y}|\mathbf{X}) - \frac{1}{2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right)p(\phi)}{\iint \exp\left(ll(\widehat{\boldsymbol{\beta}},\phi;\boldsymbol{y}|\mathbf{X}) - \frac{1}{2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right)p(\phi)\,d\boldsymbol{\beta}\,d\phi} \quad \text{Eq. 19.}$$

2. 随后，我们对分母中的 $\boldsymbol{\beta}$ 进行积分，以求得多元正态分布的归一化常数：$(2\pi)^{\frac{p}{2}}\sqrt{|\boldsymbol{\Sigma}|} = (2\pi\phi)^{\frac{p}{2}}\sqrt{\left|\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{W}_{\widehat{\boldsymbol{\beta}}}\mathbf{X}\right)^{-1}\right|}$。接下来，将分子和分母同时乘以 $\phi^{\frac{p}{2}}$：

$$p(\boldsymbol{\beta},\phi|\boldsymbol{y},\mathbf{X}) \approx \frac{\exp\left(ll(\widehat{\boldsymbol{\beta}},\phi;\boldsymbol{y}|\mathbf{X})\right)\phi^{\frac{p}{2}}p(\phi)}{\int \exp\left(ll(\widehat{\boldsymbol{\beta}},\phi;\boldsymbol{y}|\mathbf{X})\right)\phi^{\frac{p}{2}}p(\phi)\,d\phi} \times \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right)}{(2\pi)^{\frac{p}{2}}\sqrt{|\boldsymbol{\Sigma}|}} \quad \text{Eq. 20.}$$

3. 然后对后验密度中的 $\boldsymbol{\beta}$ 积分，得到 ϕ 的边际密度：

$$p(\phi|\boldsymbol{y},\mathbf{X}) \approx \frac{\exp\left(ll(\widehat{\boldsymbol{\beta}},\phi;\boldsymbol{y}|\mathbf{X})\right)\phi^{\frac{p}{2}}p(\phi)}{\int \exp\left(ll(\widehat{\boldsymbol{\beta}},\phi;\boldsymbol{y}|\mathbf{X})\right)\phi^{\frac{p}{2}}p(\phi)\,d\phi}$$ Eq. 21.

4. 最后，我们假设该边际密度在其众数 ϕ_{MAP} 附近高度集中，因此可以用狄拉克δ函数近似：$p(\phi|\boldsymbol{y},\mathbf{X}) = \delta(\phi - \phi_{MAP})$。在这种情况下，在式 18 中对 ϕ 积分，等价于令 $p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X}) \approx p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\hat{\phi})$。这一步近似得到一个经验贝叶斯估计量，作为 β 的边际后验密度的数值近似。

推导中唯一采用的近似是步骤 1 中的二阶类正态展开，以及步骤 4 中的经验贝叶斯估计量。若在中间步骤中使用广义线性模型的偏差形式和鞍点近似（Jorgensen 1987年、1997年；Nelder 和 Pregibon 1987年），那么现有的广义线性模型非贝叶斯估计方法可以解释为这一近似方案的不同变体。

- 如果对ϕ 的先验分布取均匀分布，则使用拉普拉斯近似计算公式 21 中的积分时，需要计算被积函数的众数。这直接得到修正剖面对数似然估计量$\hat{\phi}_{mPL}$。随后，经验贝叶斯估计量可由$\phi_{MAP} = \hat{\phi}_{mPL}$得到。

- 代入鞍点近似（公式 9）和针对方差参数的非信息性杰弗里斯先验$p(\phi) \propto \phi^{-1}$ 将公式 21 修改为：

$$p(\phi|\boldsymbol{y},\mathbf{X}) \approx \frac{\exp(-D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})/2\phi)\phi^{-(\frac{n-p}{2}+1)}}{\int \exp(-D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})/2\phi)\phi^{-(\frac{n-p}{2}+1)}\,d\phi}$$ Eq. 22.

该密度可识别为自由度为 $n-p$ 、尺度参数为 $D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})/(n-p)$ 的尺度逆卡方分布。因此，边缘分布 $p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X})$ 是自由度为 $n-p$ 、协方差矩阵为

$\frac{D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})}{n-p}\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{W}_{\widehat{\boldsymbol{\beta}}}\mathbf{X}\right)^{-1}$。该分布收敛于经验贝叶斯估计量，而随着有效样本量 $n-p$ 增大，该估计量又与公式 3 中的频率学派结果相同。

➢ 将鞍点近似（式 9）与均匀先验 $p(\phi) \propto 1$ 结合后，ϕ 的边际密度仍为尺度逆卡方分布，但其自由度为 $n-p-2$，尺度参数为 $D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})/(n-p-2)$。该分布的众数等于 $\phi_{MAP}=D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})/(n-p)$，因此经验贝叶斯估计值 $p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi_{MAP})$服从多元正态分布，其协方差矩阵等于 $\frac{D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})}{n-p}\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{W}_{\widehat{\boldsymbol{\beta}}}\mathbf{X}\right)^{-1}$。然而，在鞍点近似下，$p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X})$ 的精确边缘分布为自由度为 $(n-p-2)$ 的多元 t 分布。对于任意离散参数值，逆高斯分布族和高斯族的鞍点近似都是精确的，因此，在常数先验和杰弗里斯先验下得到的结果对于这些分布也是精确的（Blæsild 和 Jensen 1985；Reid 1988）。伽马密度的重归一化鞍点近似也具有这一性质（Blæsild 和 Jensen 1985；Jorgensen 1997年）。

➢ 广义线性模型的 EQL 方法是在 ϕ 的均匀先验分布下对算法 1-4 的一种修改。首先，我们将鞍点密度所隐含的对数似然代入式 17：

$$p(\boldsymbol{\beta},\phi|\boldsymbol{y},\mathbf{X})=\frac{\exp\left(-\frac{n}{2}\log(2\pi\phi)-\frac{1}{2}\sum\log(V(y_i)-\frac{1}{2\phi}\sum d(y_i,\mu_i)\right)}{\iint\exp\left(-\frac{n}{2}\log(2\pi\phi)-\frac{1}{2}\sum\log(V(y_i)-\frac{1}{2\phi}\sum d(y_i,\mu_i)\right)d\boldsymbol{\beta}\,d\phi} \qquad \text{Eq. 23.}$$

随后，我们应用拉普拉斯近似，同时处理 $\boldsymbol{\beta}$ 和 ϕ 的多重积分。由此得到联合后验分布的多元正态近似：

$$p(\boldsymbol{\beta},\phi|\boldsymbol{y},\mathbf{X})\sim MVN\left(\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\phi}_{EQL}\end{bmatrix},\begin{bmatrix}\widehat{\phi}_{EQL}\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{W}_{\widehat{\boldsymbol{\beta}}}\mathbf{X}\right)^{-1} & 0\\ 0 & \frac{2\,\widehat{\phi}_{EQL}}{n}\end{bmatrix}\right)$$

虽然与目前为止考虑的近似渐近等价，但对于小到中等有效样本量而言，这种计算方法的令人满意程度要低得多。要理解原因，请注意，这种近似等同于估计正态样本的标准差时不进行贝塞尔校正。

- EQL 的缺点可以通过采用限制最大似然（REML）估计来解决，例如参见 Lee 等人，2017 年的第 87—91 页。在 REML 方法中，首先相对于 $\boldsymbol{\beta}$ 最大化公式 23 中的对数似然；随后，在频率学派视角下得到的调整后的剖面对数似然（除一个常数外，等于关于 ϕ 的后验边际对数似然）再相对于 ϕ 进行最大化。这一系列最大化相当于一种混合的迭代拉普拉斯/符号积分过程，重现了该算法的核心：a）首先相对于 $\boldsymbol{\beta}$ 应用拉普拉斯方法（步骤 1—2），从而有效地近似 p（$\boldsymbol{\beta}$、$\phi|\boldsymbol{y}$、$\mathbf{X}$）$\approx p$（$\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}|\boldsymbol{y}$、$\mathbf{X}$、$\phi$）$p$（$\phi|\boldsymbol{y}$、$\mathbf{X}$），然后 b）对 $\boldsymbol{\beta}$ 进行解析积分（步骤 3），最后 c）再次应用拉普拉斯近似，以便为 ϕ 的后验边际分布推导出类似正态分布的近似（步骤 4）。按照这种方式计算，并在均匀先验下，ϕ 的后验边际分布的众数出现在 D（$\boldsymbol{y}$；$\widehat{\boldsymbol{\mu}}$）/（$n-p$）处。代入这一数值后，$\boldsymbol{\beta}$ 的经验贝叶斯近似后验分布与在杰弗里斯先验下为 ϕ 计算得到的结果相似。因此，换一种方式来看，REML 方法是一种分两个阶段进行的近似贝叶斯方法：在 1st 阶段，使用拉普拉斯近似并依据均匀先验对固定参数进行积分消去；而在 2nd 阶段，则获得最大似然估计值——或与其在数值上相同的贝叶斯后验众数估计。

- 矩估计量的方法也可以看作一种近似的贝叶斯计算。为此，我们将式 23 中的指数化核关于 y_i 展开至二阶。

附近μ_i。对于偏差分量，这样的操作相当于将以下项展开至二阶：

$w_i\{y_i((b')^{-1}(y_i) - (b')^{-1}(\mu_i)) - b((b')^{-1}(y_i)) + (b')^{-1}(\mu_i)\}$ 公式 24。利用恒等式 $((b')^{-1})' = 1/b'' \circ (b')^{-1}$, $b'((b')^{-1}(\mu_i)) = \mu_i$, $b''((b')^{-1}(\mu_i)) = b''(\theta_i) = V(\mu_i)$，并约去符号相反的项后，偏差的二阶展开近似等于皮尔逊残差。$V(y_i)$ 展开的首项为 $V(\mu_i)$，它可以与指数函数中的 $-\frac{n}{2}\log(2\pi\phi V(\mu_i))$ 合并；其余各项由于与 ϕ 无关，因而会在分子和分母中相互抵消，可以忽略。随后，我们可以为 ϕ 采用均匀分布先验或杰弗里斯先验，并推导出等价的后验表达式，其中偏差残差之和已被皮尔逊残差之和所替代。

总而言之，对于尺度参数已知的广义线性模型回归问题（例如泊松回归或逻辑回归），似然主义分析在数值上与对同一数据进行的近似贝叶斯分析完全相同，后者采用非信息性的均匀分布先验分布，将所有参数限制在某个有限区间内。广义线性模型各种频率学派方法的贝叶斯对应方法，也可通过对尺度参数采用两种不同的非信息先验分布来获得。对于由鞍点密度近似的广义线性模型，这些方法预计也会产生与似然主义分析几乎相同的数值结果；原则上，二者的差异可通过相应自由度下的 t 分布与正态分布之间的差异来量化。

3.4. 魔法公式、非二次似然和局部均匀先验

似然在最大值附近呈现二次行为，是贝叶斯解与频率学派解之间实现数值等价的关键要求。这也是一个可能不成立的假设，因为对于两种在生物医学领域具有重要意义的广义线性模型，即二项逻辑回归和泊松回归模型，该假设可能被违反。在

20 世纪 70 年代末和 80 年代初，一系列论文（Barndorff-Nielsen 1979；Durbin 1980；Barndorff-Nielsen 1980 年、1983 年；Hinkley 1980）研究了尺度参数已知的指数族模型中参数最大似然估计的分布表达式。这正是泊松和逻辑斯蒂广义线性模型的情形。这些研究的核心可以概括为许多作者（Ghosh 1994；Efron 1998；Pawitan 2013；Lee 等人，2017 年）此前称为似然 p-公式的内容：

$$p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)=\underbrace{\left(\frac{n}{2\pi}\right)^{\frac{p}{2}}\left|\bar{I}(\widehat{\boldsymbol{\beta}})\right|^{\frac{1}{2}}\frac{\exp\left(ll(\mathbf{X},\boldsymbol{\beta},\phi;\boldsymbol{y}|\mathbf{X})\right)}{\exp\left(ll(\widehat{\boldsymbol{\beta}},\phi;\boldsymbol{y}|\mathbf{X})\right)}}_{likelihood\ p-formula}\{1+O(n^{-1})\} \quad \text{Eq. 25.}$$

这种近似所需的元素与广义线性模型传统处理方法中构造的二次近似相同，即最大似然估计和最大值处的平均观测信息矩阵 $\bar{I}(\widehat{\boldsymbol{\beta}})=(n\boldsymbol{\Sigma})^{-1}$，然而其误差阶为 n^{-1}，而不是最大似然估计量通常具有的 $n^{-1/2}$（Barndorff-Nielsen 和Cox 1979年；Durbin1980年）。似然 p 公式的误差阶可以通过对公式25重新归一化，使其积分为1，从而得到改进。所得的 $p^{*}(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)=R(\boldsymbol{\beta})p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)$</style}近似的误差为 $n^{-3/2}$，称为Barndorff-Nielsen 的神奇 p*公式。这些公式的贝叶斯对应形式推导如下：

$$\begin{aligned}p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi)&=\frac{\exp\left(ll(\mathbf{X},\boldsymbol{\beta},\phi;\boldsymbol{y}|\mathbf{X})\right)p(\boldsymbol{\beta})}{\int\exp\left(ll(\mathbf{X},\boldsymbol{\beta},\phi;\boldsymbol{y}|\mathbf{X})\right)p(\boldsymbol{\beta})d\boldsymbol{\beta}}=\frac{\exp\left(n\bar{ll}(\mathbf{X},\boldsymbol{\beta},\phi;\boldsymbol{y}|\mathbf{X})\right)p(\boldsymbol{\beta})}{\int\exp\left(n\bar{ll}(\mathbf{X},\boldsymbol{\beta},\phi;\boldsymbol{y}|\mathbf{X})\right)p(\boldsymbol{\beta})d\boldsymbol{\beta}}\\&=\left(\frac{n}{2\pi}\right)^{\frac{p}{2}}\left|\bar{I}(\widehat{\boldsymbol{\beta}})\right|^{\frac{1}{2}}\frac{\exp\left(ll(\mathbf{X},\boldsymbol{\beta},\phi;\boldsymbol{y}|\mathbf{X})\right)}{\exp\left(ll(\widehat{\boldsymbol{\beta}},\phi;\boldsymbol{y}|\mathbf{X})\right)}\times\frac{p(\boldsymbol{\beta})}{p(\widehat{\boldsymbol{\beta}})}\{1+O(n^{-1})\}\end{aligned} \quad \text{Eq. 26.}$$

保留分母积分的拉普拉斯近似的第一项并取其倒数。在后一个表达式中，$\bar{ll}(\boldsymbol{\beta},\ \phi|\boldsymbol{y})$是统计模型的平均对数似然。所得渐近展开与 p 公式具有相同的首项，并且由于多元拉普拉斯近似的性质而同样具有$O(n^{-1})$精度（参见第

8.3 节，Bleistein and Handelsman（2010）中的公式 8.3.50）。在局部均匀先验 $\frac{p}{(\boldsymbol{\beta})}\ p(\widehat{\boldsymbol{\beta}})\approx 1$ 对所有$\boldsymbol{\beta}\ \in\ \boldsymbol{B}$成立的情况下，并且近似为 $p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi)\ =\ p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)\{1+O(n^{-1})\}$。与似然主义结果一样，我们可以对该近似进行重新归一化，使其积分为 1。重新归一化可用于将 $p(\boldsymbol{\beta})\approx p(\widehat{\boldsymbol{\beta}})$ 的误差"吸收"到归一化常数的近似误差中，从而得到具有$O(n^{-3/2})$精度的贝叶斯近似，我们将其记为$p^{*}(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi)$。

因此，重新归一化的贝叶斯近似与 p* 近似在渐近意义下的关系为：

$$\begin{aligned} p^{*}(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi) &= p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi)\{1+O(n^{-3/2})\} \\ p^{*}(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi) &= p^{*}(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)\{1+O(n^{-3/2})\} \end{aligned} \qquad \text{Eq. 27.}$$

为证明这一点，请注意，重新归一化的间接 Edgeworth 近似（Daniels 1954；Barndorff-Nielsen 和 Cox 1979）以及作为似然主义者和贝叶斯解基础的拉普拉斯近似展开具有渐近性质，这意味着存在正实常数 M_F 和 M_B：

$$p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)=\left|p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)\right|\leq p^{*}(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)\left(1+\frac{M_F}{n^{\frac{3}{2}}}\right),$$

$$p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi)=|p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi)|\leq p^{*}(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)\left(1+\frac{M_B}{n^{3/2}}\right).$$

因此：

$$\begin{aligned} \frac{p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi)}{p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)} &= \left|\frac{p^{*}(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi)}{p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)}\right| = (1+\frac{M_B}{n^{3/2}})\Big/\left(1+\frac{M_F}{n^{3/2}}\right) \\ &= 1+\frac{M_B-M_F}{n^{3/2}}\times\frac{1}{1+\frac{M_F}{n^{3/2}}} \leq 1+\frac{|M_B-M_F|}{n^{3/2}}\times\frac{1}{1+\frac{M_F}{n^{3/2}}} \\ &\leq 1+\frac{|M_B-M_F|}{n^{3/2}} \end{aligned}$$

通过类似的推理，我们可以证明：$\frac{p}{(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\ \mathbf{X},\ \phi)\ p(\boldsymbol{\beta}|\boldsymbol{y},\ \mathbf{X},\ \phi)}=\left|\frac{p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)}{p^{*}\ \boldsymbol{\beta}|\boldsymbol{y}\ \mathbf{X}\ \phi\ (\boldsymbol{,,})}\right|\leq 1+\frac{|M_B-M_F|}{n^{3\ 2}},$/从而证明似然主义解与贝叶斯解之间的$n^{-3/2}$ 渐近等价。一个实际

含义是，这种等价性意味着，用于计算$p(\boldsymbol{\beta}|\boldsymbol{y},\ \mathbf{X},\ \phi)$或由其导出的量的数值方法（例如蒙特卡洛马尔可夫链模拟），也会同时提供相应的$p^{*}(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\ \mathbf{X},\ \phi)$计算，其精度达到$n^{3/2}$ 阶；这适用于所有满足 p* 近似适用条件（具有充分统计量的模型）且采用局部均匀先验分布的模型。显然，在未进行重新归一化时，贝叶斯结果与似然主义结果相差$O(n^{-1})$，而非$O\left(n^{-3/2}\right)$。为简化讨论，我们将注意力限制在已知尺度参数的非二次情形；不过，贝叶斯结果与似然主义结果之间的联系可以直接扩展到广义线性模型鞍点近似中的干扰参数。下面概述相关步骤，不进行完整推导。首先，将边际密度的被积函数因式分解，得到：

$$p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X})=\int p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X},\phi)p(\phi|\boldsymbol{y},\mathbf{X})d\phi=\int p(\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta},\mathbf{X},\phi)p(\phi|\boldsymbol{y},\mathbf{X})\{1+O(n^{-1})\}d\phi$$

第二步，根据第3节的论证，将具有适当自由度的尺度逆卡方分布（k）代入$p(\phi|\boldsymbol{y},\mathbf{X})$。假设尺度逆卡方分布的众数远离边界，对其应用一元拉普拉斯近似，可得渐近展开$\propto p(\widehat{\boldsymbol{\beta}}\big|\boldsymbol{\beta},\mathbf{X},\hat{\phi})\{1+O(n^{-1})\}\ \{1+O\left(k^{-\frac{3}{2}}\right)\}\ \propto p(\widehat{\boldsymbol{\beta}}\big|\boldsymbol{\beta},\mathbf{X},\hat{\phi})\{1+O(n^{-1})\}\ \{1+O\left(O(n)^{-\frac{3}{2}}\right)\}\ =p(\widehat{\boldsymbol{\beta}}\big|\boldsymbol{\beta},\mathbf{X},\hat{\phi})\{1+O(n^{-1})\}$。

4. π 值，即 p 值的贝叶斯对应物

尽管此前已有研究在考察双侧p值的可重复性时考虑了后验预测分布（Goodman 1992年；Shao 和 Chow 2002年；Senn 2002），但迄今为止，尚无明显的贝叶斯依据能够为p值本身的计算提供支持。本节旨在为p值的计算和阈值化提出贝叶斯论证。为避免与频率学派的解释混淆，我们将这些数值上与p值相同的贝叶斯量称为π 值。

4.1. π 值量化方向性协变量效应的幅度

对处理效应估计值的许多解读，是围绕协变量效应的方向，而非其幅度展开的。例如，在生物科学和社会科学中，由于所涉及现象的复杂性和多样性，精确的零假设几乎总是不成立（Meehl 1967, 1978；Greenwald et al. 1996）；而且，许多理论实际上主张的是统计确定效应的方向，而非其幅度（Greenwald et al. 1996；Gelman and Tuerlinckx 2000）。在药物开发中，确认药物有效且安全是药物注册的先决条件。这两种说法都涉及方向性推断，而这些推断并不依赖于人们普遍认同的效应量或安全界限幅度。在生物医学和社会科学应用领域，测量噪声和实验单位异质性往往会限制将观察到的效应幅度（但非其方向）推广到其他情境。

基于这些考虑以及由此产生的方向性推断需求，我们现在考察广义线性模型的线性预测子中协变量效应方向（或更一般地说，与 $\beta_{i,0}$ 基线水平之差的方向）的后验边际均值。该方向由符号函数 $sgn(\beta_i - \beta_{i,0})$ 给出；当其自变量非负时取值为 1，否则取值为 −1

。在二次损失函数下，方向性的最佳点估计为后验均值：

$$\int sgn(\beta_i - \beta_{i,0})\, p(\beta_i|\boldsymbol{y}) d\beta_i = P(\beta_i \geq \beta_{i,0}|\boldsymbol{y}) - P(\beta_i < \beta_{i,0}|\boldsymbol{y}) = \quad \text{Eq. 28.}$$

$$sgn\big(P(\beta_i \geq \beta_{i,0}|\boldsymbol{y}) - P(\beta_i < \beta_{i,0}|\boldsymbol{y}, \mathbf{X})\big) \times \left[1 - \overbrace{2 \times min\big(\{P(\beta_i \geq \beta_{i,0}|\boldsymbol{y}, \mathbf{X}), P(\beta_i < \beta_{i,0}|\boldsymbol{y}, \mathbf{X})\}\big)}^{\pi - value}\right]$$

在公式 11 中，我们已证明，检验 H_0: $\beta_i = \beta_{i,0}$ 的 Wald 检验 p值，等于两个互补随机事件的概率中较小者的两倍：$\widehat{\beta_i} \geq \beta_{i,0}$ 和 $\widehat{\beta_i} < \beta_{i,0}$。其贝叶斯对应量 π 值等价地定义为后验密度中两个尾概率较小者的两倍，$\{P(\beta_i \geq \beta_{i,0}|\boldsymbol{y}, \mathbf{X}), P(\beta_i < \beta_{i,0}|\boldsymbol{y}, \mathbf{X})\}$，针对的是 β_i。π值量化了方向性协变量效应的绝对幅度，因此，其值越小，沿着 $\beta_i - \beta_{i,0}$ 符号所指方向的效应越大。总之，若使用符号函数将治疗效应映射到 $\mathrm{R} \rightarrow \{-1,0,1\}$，则通过公式 28，π值与 $\beta_{i,0}$ 左右两侧两个互补后验尾概率之差的符号，共同为二次平方损失下符号函数取值提供一个估计量。

4.2. p 值与 π 值的数值关系

对于尺度参数已知的广义线性模型，p 值与 π-值相同，因为后验分布与最大似然估计分布相同。当尺度参数必须由数据估计时，如果似然主义者（$\hat{\phi}$）和贝叶斯主义者（ϕ_{MAP}）在相同假设下计算点估计，那么似然主义者的结果与经验贝叶斯后验分布 p（$\boldsymbol{\beta}|\boldsymbol{y}$， $\boldsymbol{X}$， ϕ_{MAP}）也相同。如果对 ϕ 的不确定性取平均，从而使用 $\boldsymbol{\beta}$ 的完整正规边际后验分布，那么基于高斯分布计算的 p 值与 π 值之间的数值差异会缩小；这相当于标准正态分布的尾部概率与

相应自由度下 Student t 分布的尾部概率（St）：$n-p$ 适用于ϕ^{-1} 先验分布的情形，$n-p-2$ 适用于均匀先验分布的情形。对于高斯分布和St_{n-p}分布，区域概率对应于标准化统计量的值z_i；而对于St_{n-p-2}，相关统计量等于$z_i\sqrt{n-p/n-p-2}$。在

图 *2* 展示了在 z_i 和自由度 $n-p$ 的取值范围内，对应三个分布的（双倍）尾部概率之差。随着自由度增加，三个分布的尾部概率之间的差异（因此频率学派 p 值与后验边际分布 π 值之间的差异）实际上会消失。由于统计软件包在离散参数由数据估计时，依据高斯分布而非学生 t 分布进行 Wald 检验，因此贝叶斯 π 与频率学派 p 值之间的数值差异在实践中会很小，尤其是随着数据量增加时。此外，只要有效自由度超过三十，尺度参数先验分布的选择也不会有太大影响。

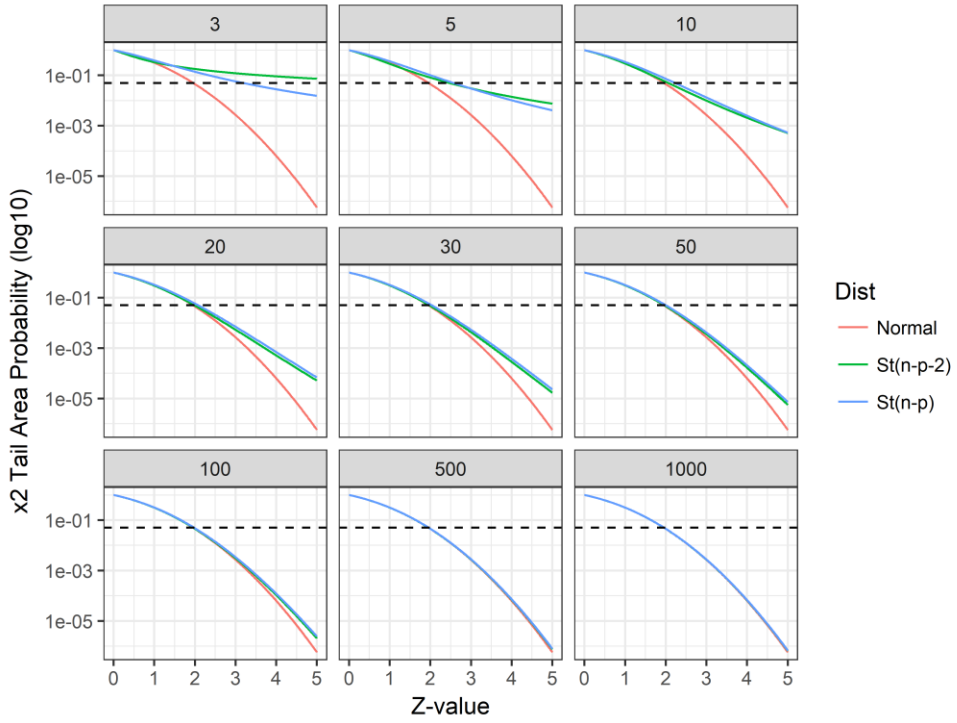


图2 标准正态分布和学生 *t* 分布（*St*）在不同自由度（各个面板）和 *Wald* 统计量（*Z* 值）下的双倍尾部概率（*π* 值，以 *log10* 为尺度）。虚线表示 *0.05* 阈值。

4.3. π 值阈值化的决策分析框架

到目前为止的论述并未要求必须对 π 值进行阈值化；事实上，贝叶斯分析师可以直接使用这些值，按照第 4.1 节的框架来量化治疗效应的幅度。本节将说明，阈值化的依据源于两个参与者之间的互动：一位分析师，负责报告广义线性模型分析结果；以及其客户。这种互动要求使用不同于公式 28 中二次损失函数的损失函数。我们将分步骤阐述这种互动：首先考察分析师对方向效应的决策（分类），然后考察客户在收到分析师的建议后必须做出的决策（Hildreth 1963）。本节还将提出一种解释，把 π 值视为分析师衡量数据价值的指标。另一方面，客户的决策则产生了对

对分析师的决策进行重新校准，以考虑科学发现的使用情境。对π值进行阈值化，是这种重新校准的自然结果。

4.3.1. "纯分析师"的决策问题

纯分析师主要关注报告真实状态，通常将其界定为估计效应与阈值 $\delta_i = \beta_i - \beta_{i,0}$ 之间的差异，并且不考虑任何下游应用。因此，纯分析师只关心作出错误陈述，其效用函数 $U_{an}(x)$ 只考虑正确陈述或错误陈述所带来的收益或损失。为了描述分析师的决策问题，我们假设分析师拥有一定的基准资本（不一定是货币形式）（C），例如声誉；正确陈述只能使其声誉提升，而错误陈述则必定会使其声誉下降。此外，分析师对世界状态没有偏好：只要分析师的判断与世界的真实状态一致，无论估计效应的方向如何，所获得的收益或损失（α）都相同。纯分析师的决策问题见表2：

具有效用函数 $U_{an}(x)$ 的纯分析师的决策表：初始资本为 C，获得数据 $\boldsymbol{y}$ 后的对称收益或损失为 α

		不确定的世界状态	
		$\delta_i \geq 0$	$\delta_i < 0$
分析师的行动	状态 $\delta_i \geq 0$	$U_{an}(C+\alpha)$	$U_{an}(C-\alpha)$
	状态 $\delta_i < 0$	$U_{an}(C-\alpha)$	$U_{an}(C+\alpha)$
概率		$P(\delta_i \geq 0 \mid \boldsymbol{y}, \mathbf{X})$	$P(\delta_i < 0 \mid \boldsymbol{y}, \mathbf{X})$

那么，分析师可能作出的两种陈述的效用为：

$$\begin{cases} U_{an}(C+\alpha) - P(\delta_i < 0 \mid \boldsymbol{y}, \mathbf{X}) \times [U_{an}(C+\alpha) - U_{an}(C-\alpha)], & \text{state } \delta_i \geq 0 \\ U_{an}(C+\alpha) - P(\delta_i \geq 0 \mid \boldsymbol{y}, \mathbf{X}) \times [U_{an}(C+\alpha) - U_{an}(C-\alpha)], & \text{state } \delta_i < 0 \end{cases}$$

让我们假设，给定数据后，两个后验尾概率分别为 $P(\delta_i \geq 0 \mid \boldsymbol{y}, \mathbf{X}) \geq P(\delta_i < 0 \mid \boldsymbol{y}, \mathbf{X}) \Longrightarrow P(\delta_i < 0 \mid \boldsymbol{y}, \mathbf{X}) \leq 0.5$。最优决策是陈述 $\delta_i \geq 0$，因为两种行动的效用之差

为 $[U_{an}(C+\alpha) - U_{an}(C-\alpha)] \times [1 - 2\,P(\delta_i < 0|\boldsymbol{y}, \mathbf{X})]$，根据假设该值为正。另一方面，如果 $P(\delta_i \geq 0|\boldsymbol{y}, \mathbf{X}) < P(\delta_i < 0|\boldsymbol{y}, \mathbf{X})$，最优决策是陈述 $\delta_i < 0$，其期望效用为 $U_{an}(C+\alpha) - P(\delta_i \geq 0|\boldsymbol{y}, \mathbf{X}) \times [U_{an}(C+\alpha) - U_{an}(C-\alpha)]$。分析师应将世界状态判定为后验尾概率最大的方向，期望效用为：

$$U_{an}(C+\alpha) - min(\{P(\delta_i \geq 0|\boldsymbol{y}, \mathbf{X}), P(\delta_i < 0|\boldsymbol{y}, \mathbf{X})\}) \times [U_{an}(C+\alpha) - U_{an}(C-\alpha)] = \quad \text{Eq. 29.}$$

$$U_{an}(C+\alpha) - \frac{[U_{an}(C+\alpha) - U_{an}(C-\alpha)]}{2} \times \pi$$

观察到数据后，分析师可能会思考这个问题的完美信息期望值（EVPI）。这是由完美信息带来的效用期望收益，这种信息能够消除做出错误决策的可能性。EVPI 可计算为不确定事件空间中各效用最大值的加权和减去 最优行动的期望效用。代入相关数值后，可得 EVPI 等于分析师效用函数在 C 处的中心差商（CDQ）、收益 α 和 π-值的乘积：

$$U_{an}(C+\alpha) - \frac{\overbrace{[U_{an}(C+\alpha) - U_{an}(C-\alpha)]}^{\text{central difference} = \delta_{2\alpha}[U_{an}](C)}}{2} \times \pi = \overbrace{\boxed{\frac{\delta_{2\alpha}[U_{an}](C)}{2\,\alpha}}}^{CDQ \approx U'_{an}(C)} \times \alpha \times \pi \qquad \text{Eq. 30.}$$

如果预期收益相对于C较小，中心差商（CDQ）就会近似等于效用函数的导数（“弹性”），因此完美信息期望值（EVPI）就等于：$U'_{an}(C) \times \alpha \times \pi - value$。由此可知，无论分析师的效用函数采用何种函数形式，完美信息期望值都与π值成正比；π值越小，完美信息对分析师而言就越没有价值。仅在这一非常特定的意义上，我们可以将π值解读为方向性效应证据强度的一种度量，因为它与完美知识的价值相关。我们必须特别强调，尽管数值相同，但在Nieman–Pearson视角下，不允许对p值作出这种解读。

4.3.2. 客户的决策问题

统计报告客户面临的决策问题，与生成该报告的分析师所面临的问题有所不同。客户收到有关某种（方向性效应）以及π值的建议后，必须决定是依据分析师对世界状态的陈述采取行动，还是对该陈述"按兵不动"。在科学研究中，这种行动可以只是对世界状态作出声明，而对分析师的陈述"按兵不动"，则可能只是以"需要进一步研究"这样平淡的措辞作总结。在监管背景下，采取行动可能意味着批准一种试验药物，而按兵不动则可能意味着延迟批准，直到收集到更多数据。与纯分析师类似，我们假设客户拥有一个效用函数$U_{co}(x)$，以及"资本"（M，限定为非负），如果客户依据分析师的陈述采取行动，那么当分析师对效应方向的判断被证明正确时，资本可以增加（"收益"），当其被证明错误时，资本则会按一定比例减少（"损失"）。另一方面，如果客户决定对分析师的陈述按兵不动，那么可能会因不行动的机会成本而遭受资本损失（表3）：

表3：统计分析客户的决策表，效用函数为 $U_{co}(x)$，初始资本为 M，收益与损失非对称，相对机会成本为 c。分数收益 (ϵ) 限定为正值，而相对损失 (ϵ') 和机会成本限定在区间 (0,1) 内

		不确定的世界状态	
		$As\ analyst\ says$	$Contrary\ to\ analyst's\ claims$
	根据分析师的陈述	$U_{co}(M \times (1+\epsilon))$	$U_{co}(M \times (1-\epsilon'))$
客户的行动	采取行动 相信分析师的 陈述	$U_{co}(M \times (1-c))$	$U_{co}(M)$
概率		$1-\frac{\pi - value}{2}$	$\frac{\pi - value}{2}$

在这种形式下，客户的决策问题过于笼统，无法清晰展现假设分数收益（ϵ）、分数损失（ϵ'）、机会成本（c）和 π 值取不同数值时所产生的行为影响。假设客户具有递减的边际效用函数，该函数可能

可以用对数函数较好地描述$U_{co}(x) = \text{对数}\, x$，那么就可以对决策问题进行数值评估。如果行动的效用大于不行动的效用，即当且仅当满足以下条件时，客户将决定根据分析师的陈述采取行动：

$$\left(1-\frac{\pi}{2}\right)\times\log\big(M\times(1+\epsilon)\big)+\frac{\pi}{2}\times\log\big(M\times(1-\epsilon')\big)>\left(1-\frac{\pi}{2}\right)\times\log\big(M\times(1-c)\big)+\frac{\pi}{2}\times\log(M)$$

因此，如果 π 值小于（大于）临界值 π_{crit}，风险厌恶型客户就会选择根据分析师的陈述采取行动（置之不理）：

$$\pi_{crit}=min\left\{1,2\times\frac{\log(1+\epsilon)-\log(1-c)}{\log(1+\epsilon)-\log(1-c)-\log(1-\epsilon')}\right\} \qquad \text{Eq. 31.}$$

临界值是收益和机会成本的非递减函数，但却是损失的非递增函数。用于判定显著性的阈值不断降低，对应于四维空间 π_{crit}、 ϵ、 ϵ'、 c 中迅速缩小的体积（图3）。公式 31及其图解解法清楚表明，阈值不能被视为静态且不可变的量，而应被视为现实世界问题的内在要素。不同的问题或不同的应用领域可能需要采用不同的π 阈值。

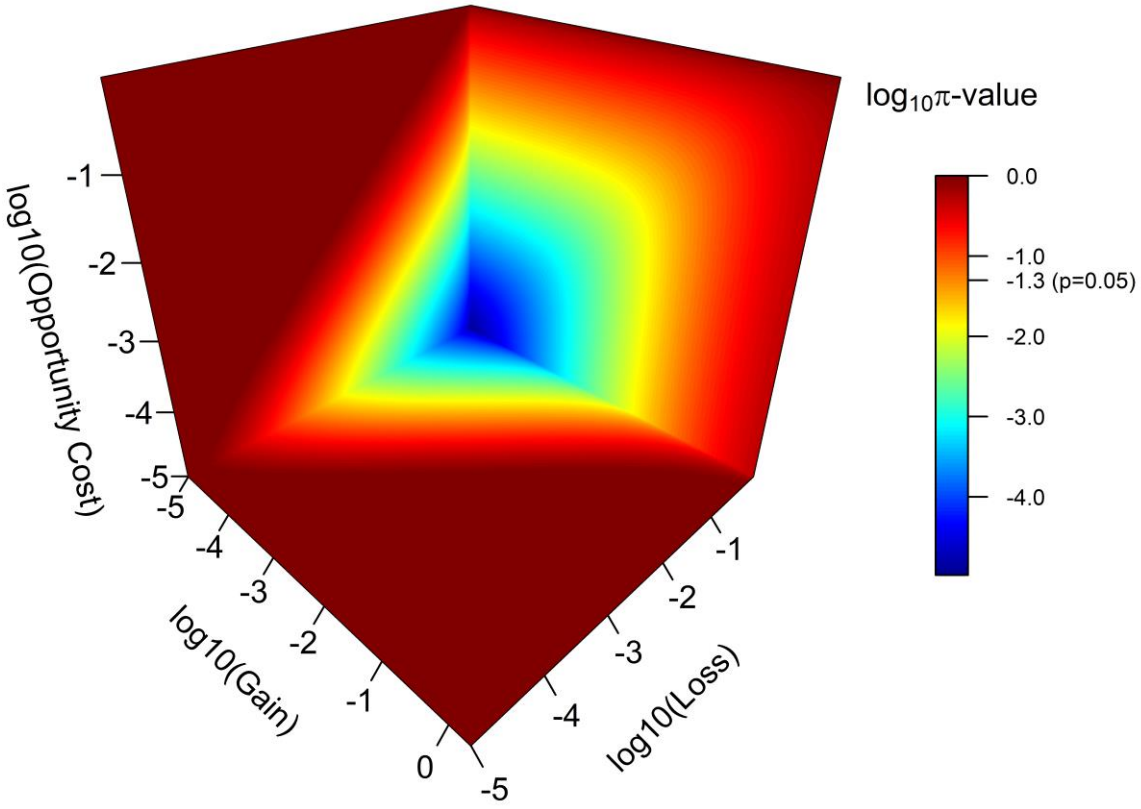


图*3* 临界阈值与分数收益、损失和机会成本之间的体积切片图。该图展示了四维空间 π_{crit}、 ϵ、 ϵ'、 c 的切片

4.3.3. “重新校准”分析师的决策问题

对于根据客户的下游行动进行“重新校准”的分析师而言，决策问题（表 4）扩展了纯分析师的选项，将关于在当前证据条件下δ_i不可采取行动性的陈述纳入其中。从本质上说，重新校准的分析者会预判客户的决策，并调整自身的效用，使其更好地与客户的目标保持一致。不可采取行动性行动对应于客户对纯分析师陈述“暂不作决定”的行动。分析师作出这一行动陈述的效用，是未进行阈值化时最优陈述效用的一部分；在给定资本C 和潜在收益α的情况下，其分数损失由函数 $f($ π_{crit}、 $\pi)$量化。

表 4 纯分析师的决策表，效用函数为 $U_{an}(x)$，初始资本为 C，对称的收益或损失为 α 在获得数据后 $\boldsymbol{y}$

		不确定的世界状态	
		$\delta_i \geq 0$	$\delta_i < 0$
分析师的行动	状态 $\delta_i \geq 0$	$U_{an}(C+\alpha)$	$U_{an}(C-\alpha)$
	状态 $\delta_i < 0$	$U_{an}(C-\alpha)$	$U_{an}(C+\alpha)$
	状态 δ_i 非-可对其采取行动当前证据	$U_{an}(C+\alpha) \times [1-f(\pi_{crit},\pi)]$	$U_{an}(C+\alpha) \times [1-f(\pi_{crit},\pi)]$
概率		$P(\delta_i \geq 0\|\boldsymbol{y},\mathbf{X})$	$P(\delta_i < 0\|\boldsymbol{y},\mathbf{X})$

我们要求，当 $\pi \leq \pi_{crit}$ 时，不采取行动声明的效用应略低于方向性声明的效用。这就是重新校准的含义，即发布一份不采取行动声明，从而建议采取客户在接受纯分析师的输入后本会自行采取的行动方案。不采取行动声明的效用被视为与正确方向性声明的效用成比例，其中比例常数是 π, π_{crit} 的函数。相应的函数 $1-f(\pi_{crit})$ 是纯分析师针对其客户将作出的下游决策的重新校准因子。对重新校准因子的可允许形式的限制，源于一项基本要求：方向性声明与不采取行动声明之间的期望效用差异，应当是差异 $\pi - \pi_{crit}$ 的递减函数 h。这一要求准确体现了以下期望目标：方向性效应的幅度越大（第 4.1 节），完美信息价值越低（第 4.3.1 节），且 π 距离行动阈值 π_{crit} 越远，不采取行动声明的效用就应越小。因此，我们要求：

$$U_{an}(C+\alpha) - \frac{\delta_{2\alpha}[U_{an}](C)}{2} \times \pi - U_{an}(C+\alpha) \times [1-f(\pi_{crit},\pi)] = \mathrm{h}(\pi-\pi_{crit}) < 0$$

满足这一要求的最简单函数是 $\mathrm{h}(\pi-\pi_{crit}) = -k \times (\pi-\pi_{crit})$，$k = k(C,a,\pi,\pi_{crit}) > 0$。因此，分数损失函数 $f(\pi_{crit},\pi)$ 是正确选择的效用的函数

方向性声明、效用函数的弹性、观测到的 π 值、π_{crit} 和比例常数 k 。如果我们选择 $k = {}^{\delta_{2\alpha}[U_{an}]}{}_{(C)\ 2}$ ， $f(\pi_{crit}, \pi)$ 就与 π 无关：

$$f(\pi_{crit}, \pi) = f(\pi_{crit}) = \frac{\delta_{2\alpha}[U_{an}](C)}{2U_{an}(C+a)} \times \pi_{crit}$$

此外，如果分析师的效用函数近似线性，那么分数损失函数就会成为正确方向性声明所带来的分数资本收益 $g = a/C$ 的简单函数：

$$f(\pi_{crit}) = \frac{g}{g+1} \times \pi_{crit} \approx \begin{cases} g \times \pi_{crit}, & g \downarrow 0 \\ \pi_{crit}, & g \uparrow \infty \end{cases}$$

这项分析表明，重新校准因子小于 1，但在数值上接近 1，因此，对于一项表明不可采取行动的声明，分析师应给予的奖励几乎应与正确方向性声明所获的奖励相同。我们注意到，对分析师而言，这种情形远比现实世界中常见的情形理想得多：分析师计算出的 p值高于显著性水平，导致客户要求分析师开展“p值搜寻”，否则就会有麻烦！最后，我们指出，重新校准的分析者的完美信息期望值变为：

$$\frac{\delta_{2\alpha}[U_{an}](C)}{2} \times min\{\pi, \pi_{crit}\} = CDQ \times \alpha \times min\{\pi, \pi_{crit}\} \qquad \text{Eq. 32.}$$

在阈值的世界里，对于π值高于阈值的分析，完美信息的价值由阈值本身决定。

5. 科学发现的重复性与P值的可重复性

科学中的重复验证危机，根源在于：当两个或更多不同的研究团队通过重复实验探究同一个问题时，已报告的科学发现无法重复验证。尽管这一术语最近再次出现（Pashler and Wagenmakers 2012），且可与特定的科学欺诈事件（Stroebe et al. 2012）以及超感现象（例如预知，Galak et al. 2012）联系起来，但事实上，人们对科学事业中不可重复性的担忧早已存在（Goodman 1992；Sterne and Smith 2001；Ioannidis 2005；Harlow et al. 2013）。重复验证某项发现的定义与显著性检验所得的p值密切相关，Goodman（Goodman 1992）对此作了精辟表述：“对许多研究者而言，重复验证意味着在一项统计功效研究中重复得到统计显著的结果。”后续关于可重复性的评论仍然聚焦于p值（Sterne and Smith 2001；Wacholder et al. 2004；Ioannidis 2005），即使作者批评了p值的使用。事实上，在分子流行病学中，“发现”往往以设定阈值的p值来表述，因此有人提出，应根据一项研究中进行的检验次数来设定p值阈值，从而控制假发现率（Benjamini and Hochberg 1995；Storey and Tibshirani 2003）。因此，尽管人们认为重复验证的含义不止于p值，实践中它却被简化为观察设定阈值的p值是否可重复。现在，我们提出一个通用的预测概率框架，用于思考重复验证问题，并可据此探究科学发现的重复验证。在这一框架中，p值的可重复性只是一个特殊且相当有限的应用场景。

5.1. 科学发现重复验证的预测概率框架

用于分析给定科学研究的重复潜力的框架，能够满足该研究结果最终使用者的需求。最终使用者可能是公众、私人机构，或最初委托、资助或要求开展该研究的官方机构。使用者并未参与研究设计（除了可能提出最初的问题，并大致说明可接受的解答方式），也没有实际开展研究或分析结果，并且对研究流程文档、测量方案和原始数据的获取程度不一。通常，使用者能够获得的唯一信息，是经过高度简化的研究问题和研究结果概要。基于这些不完整且有限的信息，使用者面临着这样的任务：调整自己的行为，以适应据称新近揭示的世界状态；投入更多资源进一步探究该问题（例如委托开展另一项研究以确认这一发现）；或者干脆转而关注其他问题。面对这些涉及机会收益与成本的选择，使用者可能会问：当前研究是否足以让人认为该研究问题已得到解决？提出这个问题时，使用者实际上将以以下信息为条件：a) 初始研究的全部数据（$\boldsymbol{y}_{init}$）和研究设计（$\boldsymbol{X}_{init}$），或 b) 足够有信息量的数据摘要，例如充分统计量（$\hat{\boldsymbol{\beta}}$）、广义线性模型拟合得到的协方差矩阵以及偏差，以便 1) 预测 重复研究的数据（$\boldsymbol{y}_{rep}$），2) 对重复数据集中的参数（$\boldsymbol{\beta}_{rep}$, ϕ_{rep}）作出推断，该数据集可能采用不同的设计矩阵（$\mathbf{X}_{rep}$），最终 3) 作出 决策。以初始数据为条件时，使用者会关心以下方面的推断：

$$p(\boldsymbol{\beta}_{rep}, \phi_{rep} | \boldsymbol{y}_{init}, \mathbf{X}_{init}, \mathbf{X}_{rep}) = \int \Bigg\{ \overbrace{\boxed{p(\boldsymbol{\beta}_{init}, \phi_{init} | \boldsymbol{y}_{init}, \mathbf{X}_{init})}}^{Inference\ from\ initial\ data} \times \overbrace{\boxed{p(\boldsymbol{\beta}_g | \boldsymbol{\beta}_{init}) p(\phi_g | \phi_{init})}}^{Translation} \times \overbrace{\boxed{p(\boldsymbol{y}_{rep} | \mathbf{X}_{rep}, \boldsymbol{\beta}_g, \phi_g)}}^{Replicate\ Data\ Generation} \times \overbrace{\boxed{p(\boldsymbol{\beta}_{rep}, \phi_{rep} | \mathbf{X}_{rep}, \boldsymbol{y}_{rep})}}^{Replicate\ Data\ Inference} \Bigg\} d\boldsymbol{\beta}_{init} d\phi_{init} d\boldsymbol{\beta}_g d\phi_g d\boldsymbol{y}_{rep}$$

Eq. 33.

在这个用于重复验证的预测框架中，我们允许将初始研究的参数翻译为生成重复研究数据的参数$(\boldsymbol{\beta}_g, \phi_g)$。在记号上，我们区分用于生成重复数据的这些参数的取值，以及利用重复数据进行推断的目标$\boldsymbol{\beta}_{rep}, \phi_{rep}$。在精确转换的设定中，$p(\boldsymbol{\beta}_g | \boldsymbol{\beta}_{init}) = \delta(\boldsymbol{\beta}_g - \boldsymbol{\beta}_{init})$且$p(\phi_g | \phi_{init}) = \delta(\phi_g - \phi_{init})$，也就是说，初始研究的参数忠实地转换为重复研究的生成过程所用参数。在精确 重复验证 $\mathbf{X}_{init} = \mathbf{X}_{rep}$ 中，重复研究的研究设计与初始研究完全相同。尽管现有研究中，可重复性的含义是精确转换和重复验证，但式 33 的框架更为一般，允许不完全迁移和非精确重复，例如样本量、研究设计或参与者特征不同的研究。通常无法通过解析方法计算式 33 中的积分，而条件模拟框架是近似求得预测分布 $p(\boldsymbol{\beta}_{rep}, \phi_{rep} | \boldsymbol{y}_{init}, \mathbf{X}_{init}, \mathbf{X}_{rep})$最直接的方法。

虽然基于公式 33 的模拟相当直接，但在以下假设下，对于精确转换和重复验证，预测分布可以推导出闭式形式：a) 广义线性模型采用鞍点近似；b) 数据似然可以用多元二次函数很好地近似

多项式，且 c) 尺度参数已知ϕ。如果有足够的数据点能够用高斯分布近似学生 t 分布，则对于相关模型，最后一个假设可以放宽。满足这些假设的模型包括泊松和逻辑斯蒂广义线性模型，以及可以用广义线性模型近似的生存模型（例如，分段指数模型或半参数比例风险模型）（Efron 1988年；Argyropoulos 和 Unruh 2015年；Bender 等，2018年）。在假设 a)-c) 下，以初始数据为条件并对重复数据进行边缘化，可替换为以充分统计量$\widehat{\boldsymbol{\beta}}_{init}$为条件并对$\widehat{\boldsymbol{\beta}}_{rep}$进行边缘化。因此，公式 33 中的积分可化简为三个多元正态分布的卷积，分别对应于：a) 对$\boldsymbol{\beta}_{init}$进行条件化的$\widehat{\boldsymbol{\beta}}_{init}$以及$\boldsymbol{\Sigma}_{init} = \phi\left(\mathbf{X}_{init}^{\boldsymbol{T}}\mathbf{W}_{\widehat{\boldsymbol{\beta}}_{init}}\mathbf{X}_{init}\right)^{-1}$的后验分布；b) 重复数据充分统计量的生成模型$\widehat{\boldsymbol{\beta}}_{rep}$；以及 c) 对$\boldsymbol{\beta}_{rep}$进行条件化的$\widehat{\boldsymbol{\beta}}_{rep}$和$\boldsymbol{\Sigma}_{rep} = \phi\left(\mathbf{X}_{init}^{\boldsymbol{T}}\mathbf{W}_{\widehat{\boldsymbol{\beta}}_{rep}}\mathbf{X}_{init}\right)^{-1}$在空间$\boldsymbol{\beta}_{init} \times \widehat{\boldsymbol{\beta}}_{rep}$上的后验分布：

$$p\left(\boldsymbol{\beta}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right) = \int p\left(\boldsymbol{\beta}_{init}\middle|\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right)p\left(\widehat{\boldsymbol{\beta}}_{rep}\middle|\mathbf{X}_{init},\boldsymbol{\beta}_{init},\phi_g\right)p\left(\boldsymbol{\beta}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{rep},\mathbf{X}_{init},\phi\right)d\boldsymbol{\beta}_{init}d\widehat{\boldsymbol{\beta}}_{rep}$$
$$= \int \underbrace{MVN\left(\widehat{\boldsymbol{\beta}}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{init},\boldsymbol{\Sigma}_{init}+\boldsymbol{\Sigma}_{rep}\right)}_{p\left(\widehat{\boldsymbol{\beta}}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right)} \times \underbrace{MVN\left(\boldsymbol{\beta}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{rep},\boldsymbol{\Sigma}_{rep}\right)}_{p\left(\boldsymbol{\beta}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{rep},\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right)} d\widehat{\boldsymbol{\beta}}_{rep}$$

由于$\boldsymbol{\Sigma}_{rep}$是$\widehat{\boldsymbol{\beta}}_{rep}$的函数，因此无法对上述表达式中的后者进行精确积分。因此，我们将解析近似代入贝叶斯定理，对$p\left(\boldsymbol{\beta}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{init},\ \mathbf{X}_{init},\ \phi\right)$进行近似：即将解析近似代入贝叶斯定理中的$p\left(\widehat{\boldsymbol{\beta}}_{rep}\middle|\boldsymbol{\beta}_{rep},\ \widehat{\boldsymbol{\beta}}_{init},\ \mathbf{X}_{init},\ \phi\right)$：

$$p\left(\boldsymbol{\beta}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right) = \frac{p\left(\boldsymbol{\beta}_{rep},\widehat{\boldsymbol{\beta}}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right)}{p\left(\widehat{\boldsymbol{\beta}}_{rep}\middle|\boldsymbol{\beta}_{rep},\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right)} \quad \text{Eq. 34.}$$
$$\approx \frac{p\left(\widehat{\boldsymbol{\beta}}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right)p\left(\boldsymbol{\beta}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{rep},\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right)}{p_{approx}\left(\widehat{\boldsymbol{\beta}}_{rep}\middle|\boldsymbol{\beta}_{rep},\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right)}$$
$$= \frac{p\left(\overline{\widehat{\boldsymbol{\beta}}}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right)p\left(\boldsymbol{\beta}_{rep}\middle|\overline{\widehat{\boldsymbol{\beta}}}_{rep},\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right)}{p_{approx}\left(\overline{\widehat{\boldsymbol{\beta}}}_{rep}\middle|\boldsymbol{\beta}_{rep},\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi\right)}$$

近似密度$p_{approx}\left(\widehat{\boldsymbol{\beta}}_{rep}\middle|\boldsymbol{\beta}_{rep}、\ \widehat{\boldsymbol{\beta}}_{init}、\ \mathbf{X}_{init}、\ \phi\right)$可通过按通常方式重新整理联合后验分布$p\left(\widehat{\boldsymbol{\beta}}_{rep}、\ \boldsymbol{\beta}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{init}、\ \mathbf{X}_{init}、\ \phi\right) = p\left(\widehat{\boldsymbol{\beta}}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{init}、\ \mathbf{X}_{init}、\ \phi\right)p\left(\boldsymbol{\beta}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{rep}、\ \widehat{\boldsymbol{\beta}}_{init}、\ \mathbf{X}_{init}、\ \phi\right)$来

通常先准备对两个高斯分布进行卷积：

$$\overbrace{MVN\left(\widehat{\boldsymbol{\beta}}_{rep}|\boldsymbol{c}(\widehat{\boldsymbol{\beta}}_{init},\boldsymbol{\beta}_{rep}),\left(\left(\boldsymbol{\Sigma}_{init}+\boldsymbol{\Sigma}_{rep}\right)^{-1}+\boldsymbol{\Sigma}_{rep}^{-1}\right)^{-1}\right)}^{p_{approx}(\widehat{\boldsymbol{\beta}}_{rep}|\boldsymbol{\beta}_{rep},\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi)}\times\overbrace{MVN(\boldsymbol{\beta}_{rep}|\widehat{\boldsymbol{\beta}}_{init},\boldsymbol{\Sigma}_{init}+2\boldsymbol{\Sigma}_{rep})}^{R(\boldsymbol{\beta}_{rep}|\widehat{\boldsymbol{\beta}}_{rep},\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi)}$$

$$\boldsymbol{c}(\widehat{\boldsymbol{\beta}}_{init},\boldsymbol{\beta}_{rep})=\left(\left(\boldsymbol{\Sigma}_{init}+\boldsymbol{\Sigma}_{rep}\right)^{-1}+\boldsymbol{\Sigma}_{rep}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_{rep}^{-1}\boldsymbol{\beta}_{rep}+\left(\boldsymbol{\Sigma}_{init}+\boldsymbol{\Sigma}_{rep}\right)^{-1}\widehat{\boldsymbol{\beta}}_{init}\right)$$

上式左侧标记为 $p_{approx}(\widehat{\boldsymbol{\beta}}_{rep}|\boldsymbol{\beta}_{rep},\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi)$ 的因素是一个近似，因为右侧的余项因素与 $\widehat{\boldsymbol{\beta}}_{rep}$ 有关，这是因为矩阵 $\boldsymbol{\Sigma}_{rep}$ 是该量的函数。

为了完成式 34 的推导，我们必须运用启发式论证来为 $\bar{\bar{\boldsymbol{\beta}}}_{rep}$ 选择一个值，以确定近似：首先，我们应用拉普拉斯近似，对$p_{approx}(\widehat{\boldsymbol{\beta}}_{rep}|\boldsymbol{\beta}_{rep},\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi)\times R(\boldsymbol{\beta}_{rep}|\widehat{\boldsymbol{\beta}}_{rep},\widehat{\boldsymbol{\beta}}_{init},\mathbf{X}_{init},\phi)$关于 $\boldsymbol{\beta}_{rep}$ 进行积分，方法是选取使余项最大化的后一个变量的值，即 $\boldsymbol{\beta}_{rep}=\widehat{\boldsymbol{\beta}}_{init}$。积分后，我们得到边际密度 $MVN\left(\widehat{\boldsymbol{\beta}}_{rep}\middle|\widehat{\boldsymbol{\beta}}_{init},\left(\left(\boldsymbol{\Sigma}_{init}+\boldsymbol{\Sigma}_{rep}\right)^{-1}+\boldsymbol{\Sigma}_{rep}^{-1}\right)_{-1}\right)$。该密度的近似众数位于 $\widehat{\boldsymbol{\beta}}_{init}$，因此这是 $\widehat{\boldsymbol{\beta}}_{rep}$最可能的值。由于我们希望式 34 中的近似在 $\widehat{\boldsymbol{\beta}}_{rep}$最可能的值附近尽可能准确，因此在式 34 中令 $\bar{\bar{\boldsymbol{\beta}}}_{rep}=\widehat{\boldsymbol{\beta}}_{init}$ ，并对行列式平方根的乘积进行化简、约去公项后，得到：

$p(\boldsymbol{\beta}_{rep},\ |\boldsymbol{y}_{init},\ \mathbf{X}_{init},\ \phi)=p(\boldsymbol{\beta}_{rep},\ |\widehat{\boldsymbol{\beta}}_{init},\ \mathbf{X}_{init},\ \phi)\approx\ MVN(\widehat{\boldsymbol{\beta}}_{init},\ 3\times\boldsymbol{\Sigma}_{init})$ 公式35。因此，在尺度参数已知且满足精确转换和重复验证的条件下，重复研究中推断的预测密度是一个多元正态分布，其中心位于本研究的点估计值，但其协方差矩阵是初始研究中获得的协变量矩阵的一个标量倍数。通过计算π值，并对其进行第 4 节所述类型的阈值分析，可以利用这一预测密度预判重复研究中作出的决策。精确重复验证的预测π值与

观测到的π值根据以下公式与初始研究的$\pi_{rep} = 2\Phi\left(\Phi^{-1}(\pi_{init}/2)\ /\sqrt{3}\right)$相关。使用 0.05 作为 p 和π值的阈值，我们可以看到，只有初始 p 值不超过 0.0006 时，预测π值才低于该阈值。

虽然式 33 假设初始数据和重复数据都将从贝叶斯视角进行分析，但稍作相应调整后，它也允许在非贝叶斯框架下分析可重复性。从非贝叶斯的角度看，重复数据集中的推断所用的预测分布由$p(\widehat{\boldsymbol{\beta}}_{rep}|\mathbf{X}_{init}, \mathbf{X}_{rep}, \boldsymbol{\beta}_{init}, \hat{\phi}_{init}, \boldsymbol{\beta}_{rep},\ \hat{\phi}_{rep})$给出，而翻译和多元积分现在都涉及$\widehat{\boldsymbol{\beta}}_{init}$ ，而非$\boldsymbol{\beta}_{init}$。乘积$p(\boldsymbol{\beta}_g|\widehat{\boldsymbol{\beta}}_{init})p(\phi_g|\hat{\phi}_{init})$提供了一种机制，可将初始研究中的估计量转换为重复数据生成模型的参数。与贝叶斯情形一样，将这两个概率密度函数分别设为狄拉克δ函数，就对应于精确转换。根据第2节和第3节所述内容，在（局部）均匀先验$p(\boldsymbol{\beta}_{init}, \phi_{init}|\boldsymbol{y}_{init}, \mathbf{X}_{init}) \approx p(\widehat{\boldsymbol{\beta}}_{init}, |\mathbf{X}_{init}, \boldsymbol{\beta}_{init}, \hat{\phi}_{init})$和$p(\boldsymbol{\beta}_{rep}, \phi_{rep}|\mathbf{X}_{rep}, \boldsymbol{y}_{rep}) \approx p(\widehat{\boldsymbol{\beta}}_{rep}|\mathbf{X}_{rep}, \boldsymbol{y}_{rep}, \boldsymbol{\beta}_{rep}, \hat{\phi}_{rep})$下，对于二次似然且尺度参数已知的情况，两者完全相等。因此，鉴于这两种不同统计思想学派用于推断的概率在数值上高度一致，式 33 以及式 34 和 35 中所述的计算可用于在贝叶斯或非贝叶斯预测概率框架下分析重复验证。

5.2. P 和 π– 值的可重复性较差

第5.1节的预测框架利用重复数据和重复参数的联合预测分布。另一种方法是采用边际预测框架，在该框架中，将每个重复数据集的参数后验概率用作下游应用的输入，例如计算π值。所有关于p值可重复性概率的既往（贝叶斯）分析都隐含地采用了这种边际框架（Goodman 1992年；Shao 和 Chow 2002年）。下面我们将介绍一种

基于模拟的方法，用于在该框架下分析p值和π值，然后给出尺度参数已知或估计得非常精确的模型的近似公式。这种层次模拟策略将用于分析第 6 节所考虑的临床试验数据集，并通过从下方示意图所示的相关"精确"（由MCMC生成的）分布或近似分布中抽样，为检验本文所采用的鞍点近似和二次似然近似提供一种直接的方法：

$$
\begin{aligned}
\boldsymbol{\beta}_{init}, \phi_{init} | \boldsymbol{y}_{init}, \mathbf{X}_{init} &\sim p(\boldsymbol{\beta}_{init}, \phi_{init} | \boldsymbol{y}_{init}, \mathbf{X}_{init}) \approx \\
&MVN\left(\boldsymbol{\beta}_{init} | \widehat{\boldsymbol{\beta}}_{init}, \boldsymbol{\Sigma}_{init}\right) \times \text{Inv-}\chi^2\left(n-p, \frac{D\left(\boldsymbol{y}_{init}; \widehat{\boldsymbol{\mu}}_{init}\right)}{n-p}\right) \\
\boldsymbol{\beta}_g, \phi_g | \boldsymbol{\beta}_{init}, \phi_{init} &\sim p(\boldsymbol{\beta}_g | \boldsymbol{\beta}_{init}) p(\phi_g | \phi_{init}) \\
\boldsymbol{y}_{rep} | \mathbf{X}_{rep}, \boldsymbol{\beta}_g, \phi_g &\sim f(\boldsymbol{y}_{rep} | \mathbf{X}_{rep}, \boldsymbol{\beta}_g, \phi_g) \\
\boldsymbol{\beta}_{rep}, \phi_{rep} | \mathbf{X}_{rep}, \boldsymbol{y}_{rep} &\sim p\left(\boldsymbol{\beta}_{rep}, \phi_{rep} | \mathbf{X}_{rep}, \boldsymbol{y}_{rep}\right) \\
&\approx MVN\left(\boldsymbol{\beta}_{rep} | \widehat{\boldsymbol{\beta}}_{rep}, \boldsymbol{\Sigma}_{rep}\right) \times \text{Inv-}\chi^2\left(n-p, \frac{D\left(\boldsymbol{y}_{rep}; \widehat{\boldsymbol{\mu}}_{rep}\right)}{n-p}\right) \\
\pi_{rep,i} &= 2 \times min\left(\left\{P(\beta_{rep,i} \geq \beta_{i,0} | \boldsymbol{y}_{rep}, \mathbf{X}_{rep}), P(\beta_{rep,i} < \beta_{i,0} | \boldsymbol{y}_{rep}, \mathbf{X}_{rep})\right\}\right) \\
&\approx p_{rep,i} \\
p_{rep,i} &= 2 \times \Phi\left(-\frac{\left|\hat{\beta}_{rep,i} - \beta_{i,0}\right|}{\sqrt{\boldsymbol{\Sigma}_{rep}[i,i]}}\right) = 2 \times \Phi\left(-\left|\hat{z}_{rep,i}\right|\right)
\end{aligned}
\qquad \text{Eq. 36.}
$$

在第 5.2 节所考虑的精确转换和重复验证设置下，重复数据集中 p 值和 π 值分布的解析近似，可先计算后验边际分布（关于 $\boldsymbol{\beta}_{init}$）$\widehat{\boldsymbol{\beta}}_{rep} | \mathbf{X}_{rep}, \boldsymbol{y}_{rep}, \phi_{rep} \sim MVN(\widehat{\boldsymbol{\beta}}_{init}, 2\boldsymbol{\Sigma}_{init})$，再对除第 i^{th} 个分量外的所有分量积分消去，从而得到：$\hat{\beta}_{rep,i} \sim N(\hat{\beta}_{rep,i}, \sqrt{2 \times \boldsymbol{\Sigma}_{rep}[i,i]})$。相应检验统计量的分布为 $\hat{z}_{rep,i} \sim N(\hat{z}_{init,i}, \sqrt{2})$。对 $\hat{z}_{rep,i} \rightarrow \pi_{rep,i}$ 进行直接的随机变量变换，便可得到 π 值的分布。先前的研究（Lambert 和 Hall 1982年；Shao 和 Chow 2002年；De Martini 2008年；Boos 和 Stefanski 2011年；De Capitani 和 De Martini 2016年）表明，考察 p 值及其值的变异性时，自然的尺度是对数刻度。此外，十进制系统表明，对数刻度应采用以 10 为底的对数；在该尺度下，$\pi_{rep,i} = x \cdot 10^{-k} \rightarrow -\log_{10} \pi_{rep,i} = k - \log_{10} x$。因此：

$$-\log_{10} p_{rep,i} = -\log_{10} \pi_{rep,i} \sim 10^{-\log_{10} \pi_{rep,i}} \exp\left(\mathrm{erfc}^{-1}\left(10^{-\log_{10} \pi_{rep,i}}\right)^2\right) \log 10 \sqrt{\frac{\pi}{2}} \times \left[N\left(\Phi^{-1}\left(\frac{10^{-\log_{10} \pi_{rep,i}}}{2}\right) \middle| \Phi^{-1}\left(\frac{10^{-\log_{10} \pi_{init,i}}}{2}\right), \sqrt{2}\right) + N\left(-\Phi^{-1}\left(\frac{10^{-\log_{10} \pi_{rep,i}}}{2}\right) \middle| \Phi^{-1}\left(\frac{10^{-\log_{10} \pi_{init,i}}}{2}\right), \sqrt{2}\right)\right]$$

公式 37。

公式 37 中重复 p 值在对数尺度和未变换尺度下的期望，与第 5.2 节的预测性 p 值在图 3A 中进行了比较。预测方法对重复概率的评估更为保守。公式 37 的 RPD 绘于图 3B：随着初始 p 值变小（$-\log_{10} \pi_{rep,\ i}$ 变得更为正向），分布逐渐转变为偏斜的“对数正态”形状。即使达到“对数正态性”，相关密度仍相当分散，例如，当 p 值为 = 0.00001 时，重复得到的 p 值中仍有相当一部分为 > 0.05。在图 3C 和图 3D 中，我们绘出了标准差与均值在以 log10 为尺度或未变换尺度下的关系。我们观察到，RPD 的标准差在 log10 尺度下占均值的相当大比例，在未变换尺度下则大于均值。RPD 的大方差清楚地表明，P 值和 π 值作为随机变量具有过大的变异性；即使初始 p 值似乎表明研究结果应能重复验证，其数值也难以重复获得。下一节将展示如何运用这些概念分析临床试验的可重复性；附录提供了说明相关分析编程的 R 代码。

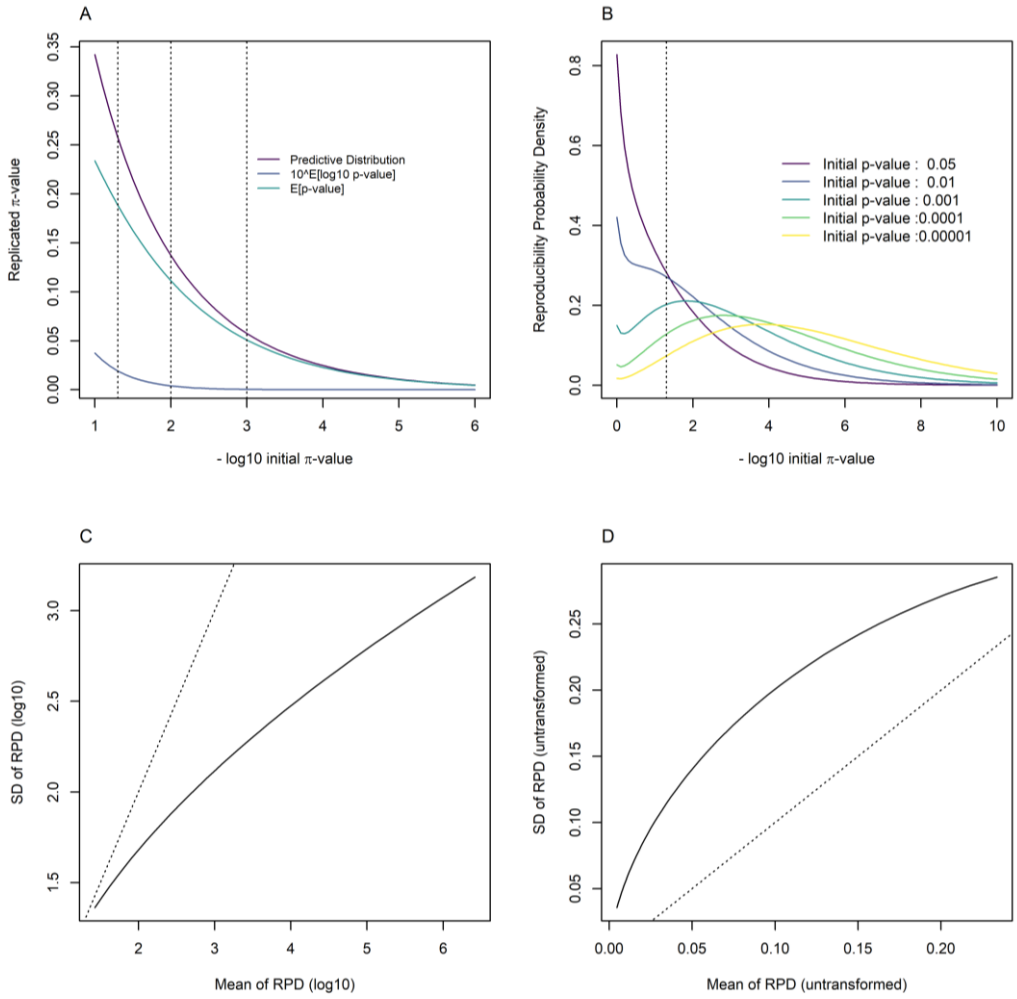


图 *4*（*A*）重复数据集中的预测 *π* 值与初始实验中的 *π* 值之间的关系（以 *-log10* 为尺度），展示预测分布以及以 *log10* 尺度和未变换尺度计算的 *RPD* 期望值。自左向右的垂直线分别对应传统的统计显著性水平 *0.05*、*0.01*、*0.001*。（*B*）不同初始 *p* 值对应的重复概率密度（垂直线对应 *p=0.05*）。（*C*）以 *log10* 为尺度的 *RPD* 标准差与均值。（*D*）未变换尺度下的 *RPD* 标准差与均值。在（*C*）和（*D*）中，点线为恒等线

6. SGLT2i肾脏结局临床试验中的可重复性

钠-葡萄糖协同转运蛋白2抑制剂（钠-葡萄糖协同转运蛋白2抑制剂）最初被作为一类降糖效果温和的药物引入，可使新接受治疗的2型糖尿病患者的糖化血红蛋白A1c降低0.81至1.02%，在二甲双胍治疗基础上加用时则可降低0.57至0.63%（Tsapas等，2020年）。这些药物的靶点是肾脏内表达的一种蛋白质（钠-葡萄糖协同转运蛋白2），其正常功能是重新吸收肾小球（即肾脏内微小的滤过单位）非选择性滤过的葡萄糖分子，并将葡萄糖与钠一起送回血液循环。糖尿病患者体内循环的葡萄糖过多，导致钠-葡萄糖协同转运蛋白2系统过度激活，肾脏滤过功能发生病理性增强；随着肾脏内的肾小球逐渐耗竭，这一过程会引发慢性肾脏病（慢性肾脏病）。过去二十年半间开展的实验涉及天然存在的钠-葡萄糖协同转运蛋白2抑制剂（根皮苷（Vallon等，1999年））以及缺乏钠-葡萄糖协同转运蛋白2的糖尿病小鼠（Vallon等，2012年），这些实验提示钠-葡萄糖协同转运蛋白2抑制剂实际上可能有助于延缓慢性肾脏病的进展。为检验特定钠-葡萄糖协同转运蛋白2抑制剂对慢性肾脏病的影响，开展了两项人体研究：一项是针对2型糖尿病患者的卡格列净 CREDENCE 试验（Perkovic等，2019年），另一项是达格列净 DAPA-CKD 试验（Heerspink等，2020年），纳入了患有和未患有2型糖尿病的患者。两项研究的主要结局均为复合终点，即至以下事件发生的时间：a）肾功能较基线下降（估计肾小球滤过率，eGFR（参考范围90-120 ml/min/1.73m2），DAPA-CKD中下降50%，或DAPA-CKD中血清肌酐翻倍）；b）出现终末期肾病（需要透析、移植或 eGFR < 15 ml/min/1.73m2）；或 c）死于心血管或肾脏原因。这两项试验中主要结局和一种罕见副作用（糖尿病酮症酸中毒）的结果汇总于表5。我们还在该表中根据这些试验的汇总数据拟合泊松广义线性模型，计算了相对风险。泊松模型估计值与主要分析中使用的比例风险模型所得结果几乎完全一致。

表 *5* 钠-葡萄糖协同转运蛋白*2*抑制剂（*SGLT2i*）肾脏结局试验 *CREDENCE*（卡格列净）和 *DAPA-CKD*（达格列净）中的主要结局和糖尿病酮症酸中毒（*DKA*）事件。表中显示试验报告的风险比、事件数和事件发生率（*DAPA-CKD* 试验未报告 *DKA*结局的事件发生率）

				钠-葡萄糖协同转运蛋白2抑制剂	安慰剂	钠-葡萄糖协同转运蛋白2抑制剂	安慰剂
结局	试验	相对风险*	风险比	N / 总人数	N / 总人数	事件数/ 1000 PYs	事件数/ 1000 PYs
主要 (1o)	CREDENCE	0.71 (0.60 – 0.83)	0.70 (0.59 – 0.82)	245/2202	340/2199	43.2	61.2
DKA	CREDENCE	7.79 (1.48 – 41.09)	10.8 (1.39 – 83.65)	11/2200	1/2197	2.2	0.2
主要 (1o)	DAPA-CKD	0.61 (0.51 – 0.73)	0.61 (0.51 – 0.72)	197/2152	312/2152	46	75
DKA	DAPA-CKD	0.00 (0.00 – ∞)	–	0/2149	2/2149	–	–

* 作者使用泊松回归模型计算，偏置项采用：*a*）研究者报告的事件数/事件发生率所隐含的随访时间（*CREDENCE/DAPA-CKD* 的主要结局以及 *CREDENCE* 的 *DKA*）；*b*）各研究组的规模（*DAPA-CKD* 的 *DKA*）。后一个回归模型在对数尺度上的点估计值和标准误为 -23（42247.17），由此得出精确点估计值 1。03 10^{-10}，95% 置信区间为 3。16 10^{-35972} - 3。36 10^{35951} 。PY：人年

图5显示了泊松回归系数（对数相对风险）的最大似然估计周围的对数似然等高线。对于 CREDENCE 和 DAPA-CKD 中的主要结局，这些等高线在平面上呈椭球形，至少在每个坐标轴的4个标准误范围内如此。对于糖尿病酮症酸中毒结局，则呈现出不同的模式：a) 在 CREDENCE 中，椭球形模式并不完全成立：沿着从左上角延伸至右下角的椭球主轴，其上方和下方的概率等概率线之间的距离不同；b) 在 DAPA-CKD 中，等概率线完全不像椭圆形：事实上，对于分配至 SGLT2抑制剂组的受试者，似乎在相对风险为1.0处存在明显的不连续性。将实际似然以三维形式可视化（图6）时，我们看到，任一研究的1o结局似然都呈现出熟悉的二元高斯密度形态。CREDENCE 中糖尿病酮症酸中毒结局的似然明显偏离二元高斯形态，因为图的右下角可见一个非对称尾部。最后，CREDENCE 中糖尿病酮症酸中毒结局的似然看起来像是将一元高斯分布“拼接”在一起。根据第2节和第3节所述内容，我们会预期二次近似对于

对两项研究中主要结局的分析均如此，但对 CREDENCE 中的 DKA 结局而言，这种情况较少见；而在 DAPA-CKD 中的 DKA 结局分析中则完全没有这种情况。

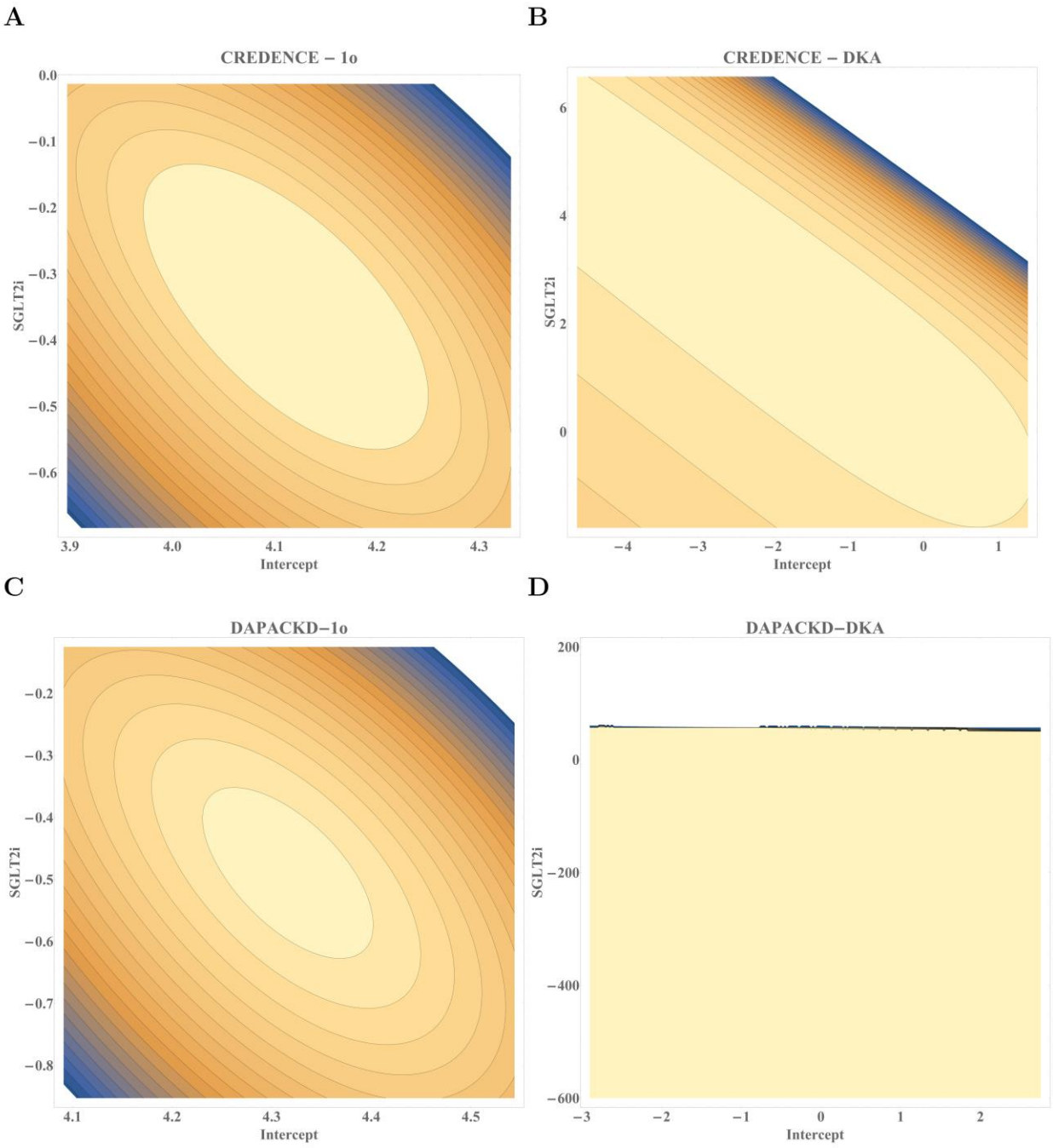


图*5 CREDENCE* 和 *DAPA-CKD* 试验中主要（*1o*）结局及糖尿病酮症酸中毒（*DKA*）的对数似然等高线。*X*轴：泊松模型的截距；*Y*轴：*SGLT2*抑制剂治疗的治疗效应

推论是：对于两项研究中的主要结局，贝叶斯分析都不会受到先验分布选择的影响。然而，对于 CREDENCE，DKA 的分析会部分受到先验分布选择的影响；而对于 DAPA-CKD，若要得到比整个正实数集更有用的结果，则完全取决于先验分布。

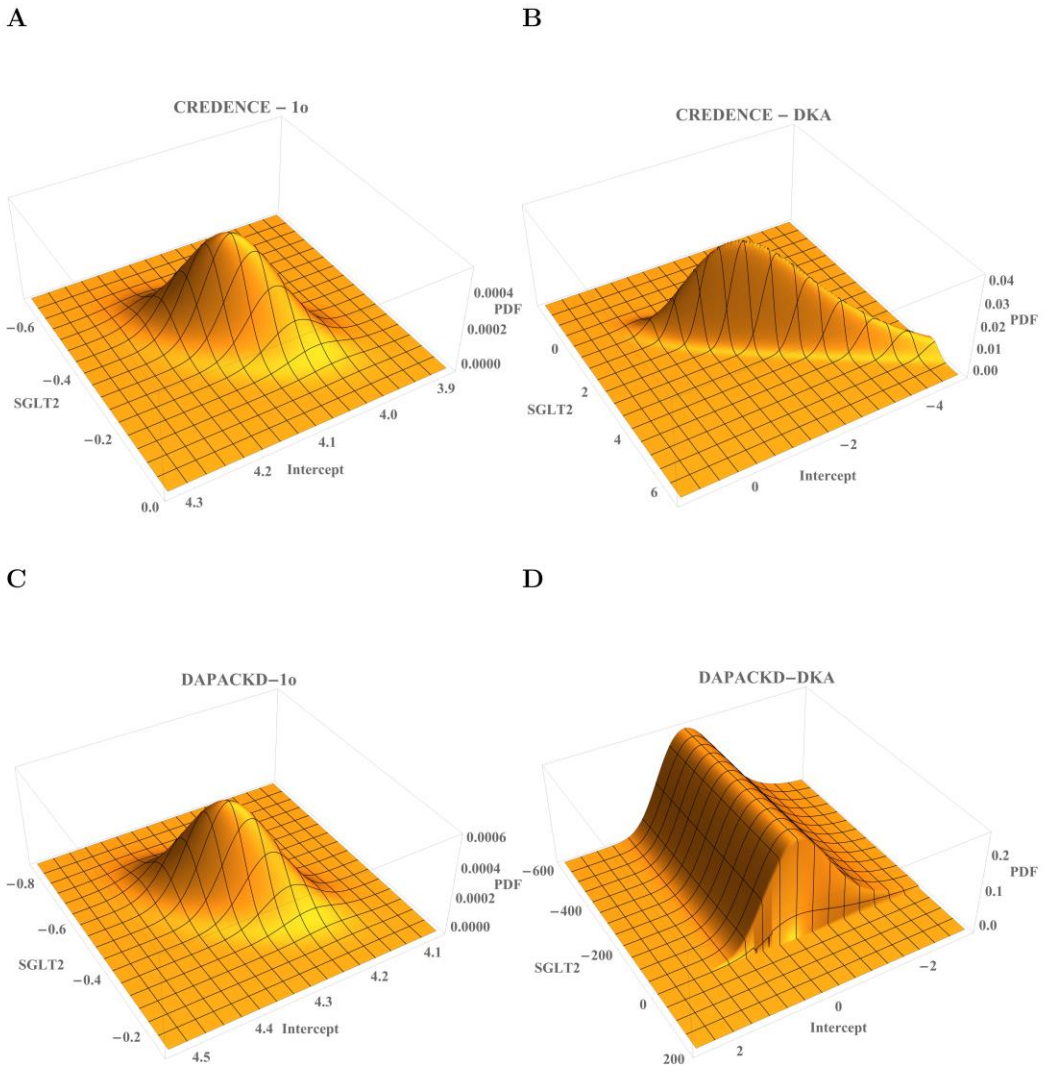


图*6 CREDENCE* 和 *DAPA-CKD* 试验中主要（*1o*）结局和糖尿病酮症酸中毒（*DKA*）的似然。*X* 轴：泊松模型的截距；*Y* 轴：*SGLT2*抑制剂治疗的处理

对于贝叶斯重新分析，我们采用了常数先验分布，或由独立柯西分布（截距）和自由度为 2.5、效应尺度参数为 1.0 的学生t分布构成的信息性集合

钠-葡萄糖协同转运蛋白2抑制剂。 后一种先验分布比 为广义线性模型引入的自由度为 1、尺度参数为 2.5 的“弱信息”学生t分布先验分布（Gelman等，2008年）更具信息性（弥散程度更低），但其信息性又不及该论文中用于截距的标准柯西分布（名义上是自由度为 1、尺度为 1 的学生t分布）。随后，我们通过 R 包“NUTS采样器”的实现，使用“rstanarm”开展贝叶斯分析（Goodrich等，2020年）。我们还使用 R 包——“arm”（Gelman等，2008年）中实现的近似期望最大化（EM）算法，在学生t先验分布下开展了贝叶斯分析。对于基于马尔可夫链蒙特卡罗的分析，我们模拟了四条相互独立的链；采样器生成的前 10,000 个样本作为预热期丢弃，随后迭代 10,000 次。为计算 π值，我们通过 EM 算法拟合有限高斯混合分布，对 MCMC样本进行平滑处理，然后使用估计的混合比例以及各成分的均值和标准差计算尾部概率。最大似然和贝叶斯估计值（以后验均值和后验标准差汇总）见表6和图7。正如下文所述，常数先验分布会导致 DAPA-CKD 中糖尿病酮症酸中毒结局的后验分布不适定，因此改用弥散高斯先验分布。

表*6*：钠-葡萄糖协同转运蛋白*2*抑制剂（*SGLT2i*）肾脏结局试验 *CREDENCE*（卡格列净）和 *DAPA-CKD*（达格列净）的主要结局和糖尿病酮症酸中毒（*DKA*）分析。表中列出了最大似然法的点估计值及其标准误（对于二次对数似然，这也近似于后验分布），以及使用 *NUTS*采样器估计平坦先验和学生*t*信息先验下的贝叶斯分析结果。此外，还列出了基于期望最大化（*EM*）的学生*t*先验近似贝叶斯分析结果。

结局	试验	最大值似然	Flat 先验分布（NUTS）	学生t分布 先验分布（期望最大化）	学生t分布 先验分布（NUTS）
主要 (1o)	CREDENCE	-0.3483 (0.0838)	-0.3483 (0.0835)	-0.3478 (0.0837)	-0.3449 (0.0834)
DKA	CREDENCE	2.3990 (1.0440)	2.9159 (1.3047)	1.9633 (0.8084)	1.7433 (0.8697)
主要 (1o)	DAPA-CKD	-0.4889 (0.0910)	-0.4896 (0.0904)	-0.4880 (0.0909)	-0.4831 (0.0912)
DKA	DAPA-CKD	-22.9957 (42247.1657)	-71.3650 (54.2870)	-1.2011 (1.1782)	-1.0364 (1.4168)

对于主要结局的分析，在任一研究中，对数似然的二次行为都会占据主导，其相对尖锐程度会盖过贝叶斯与非贝叶斯估计方法、EM与马尔可夫链蒙特卡罗算法之间的计算差异，以及先验分布本身所造成的任何推断差异：点估计值和标准误精确到小数点后三位均完全相同。对于 CREDENCE 试验中的 DKA结局，对数似然偏离二次形式的程度适中，因此最大似然和期望最大化的估计值与马尔可夫链蒙特卡罗得出的估计值有所不同。偏差最大的是 CREDENCE 试验中的 DKA结局：最大似然分析和平坦先验下的马尔可夫链蒙特卡罗分析都得出了不合理的结果。学生t先验有效地对该分析进行了正则化，得到的后验分布估计值较为合理，但估计精度较低。

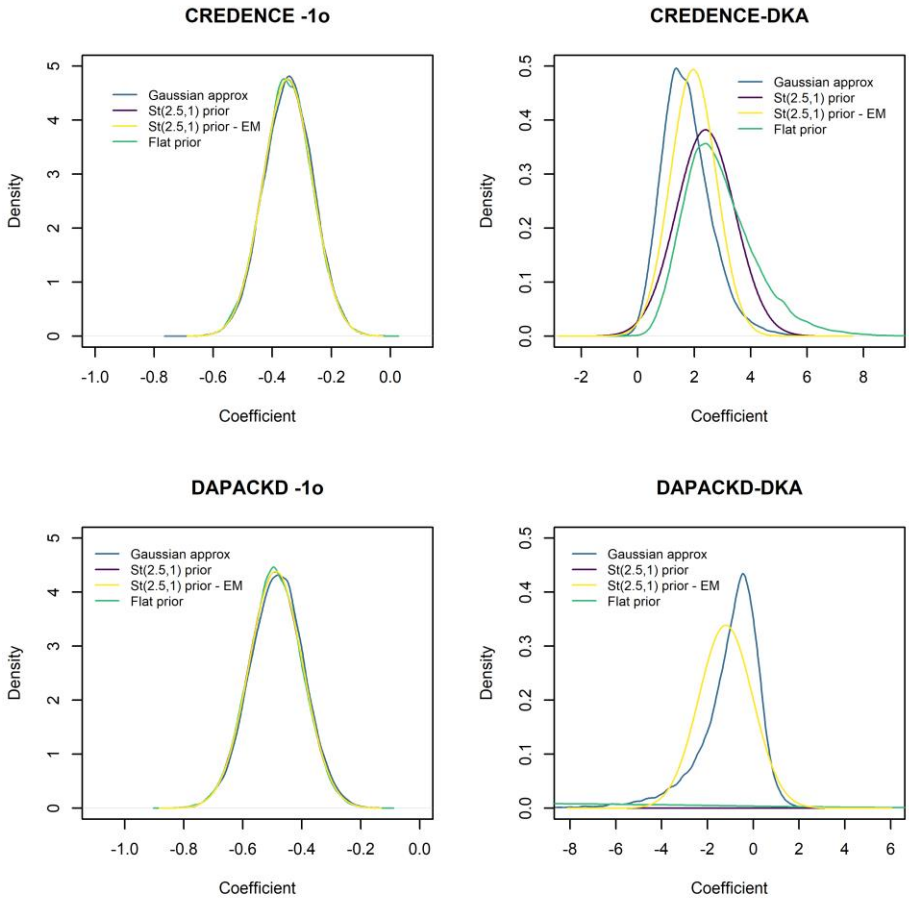


图*7 CREDENCE* 和 *DAPA-CKD* 试验中，钠-葡萄糖协同转运蛋白*2*抑制剂对肾脏特异性主要结局和糖尿病酮症酸中毒（*DKA*）的治疗效应后验概率。高斯近似基于第*2*至*3*节的材料，并与最大似然分布一致；平坦先验 = 在常数先验分布下进行马尔可夫链蒙特卡罗分析；*EM*=期望最大化。

P 和 π值见表7，说明这些值对所采用的具体先验分布以及执行贝叶斯计算的方法并不敏感。一个值得注意的分歧是，在平坦先验下，CREDENCE 中 DKA结局的π值；根据 MCMC 链计算出的π值为 0.015，这反映出仅有少量样本估计的对数风险比大于 0，而根据二次近似计算出的π值为 >0.999。

表*7* 钠-葡萄糖协同转运蛋白*2*抑制剂（*SGLT2i*）的肾脏结局试验 *CREDENCE*（卡格列净）和 *DAPA-CKD*（达格列净）中主要结局和糖尿病酮症酸中毒（*DKA*）的分析。表中显示了采用最大似然法得到的 *π*-值（也近似等于 *π*-值），以及使用 *NUTS* 采样器估计的贝叶斯分析在平坦先验分布和学生*t*信息先验下得到的 *π*-值。此外，还展示了基于期望最大化（*EM*）方法针对 *Student-t* 先验分布进行的近似贝叶斯分析结果。

结局	试验	最大值似然	Flat 先验分布 (NUTS)	学生t分布 先验分布（期望最大化	学生t分布 先验分布（NUTS）
主要 (1o)	CREDENCE	$3.23\ 10^{-5}$	$3.05\ 10^{-5}$	$3.28\ 10^{-5}$	$3.5\ 10^{-5}$
DKA	CREDENCE	0.022	0.002	0.015	0.016
主要 (1o)	DAPA-CKD	$7.79\ 10^{-8}$	$6.16\ 10^{-8}$	$7.97\ 10^{-8}$	$11.7\ 10^{-8}$
DKA	DAPA-CKD	0.999	0.015	0.308	0.343

为探究这些差异，我们针对泊松广义线性模型的截距参数，对 DAPA-CKD 中 DKA结局的似然进行了解析积分，得到函数 $\left(1+e^{\beta_1}\right)^{-2}$。对$\beta_1$（达格列净的治疗效应）积分将得到后验分布的归一化常数；然而，积分 $\int\left(_{-\infty}^{\infty}\ 1+e^{\beta_1}\ ^{-2}d\beta_1\ \right)$发散，因此后验分布是不适定的。引入学生 t 分布先验后，似然× 先验的积分不再发散，单变量数值积分得到的π值为 0.334，与表7 中 DAPA-CKD 的 DKA 对应条目相吻合。然而，这些π值的具体数值取决于学生 t 分布的尺度，并且可以通过选择尺度值使其任意小（未显示）：当数据量有限时，贝叶斯后验分布会受到

先验分布，而非贝叶斯分析以及采用非正则化先验的贝叶斯分析将得出毫无意义的数值。

参见第5.1节，CREDENCE 和 DAPA-CKD 主要结局的预测π值分别为 0.0164 和 0.0019，均低于 0.05 的名义阈值。我们还依据第 5.2 节的边际预测框架，采用公式36中的模拟方案，对重复验证进行了分析。具体而言，我们在平坦先验下进行了贝叶斯分析，并通过马尔可夫链蒙特卡罗从后验分布中生成样本。随后，我们利用每个后验样本生成泊松分布计数，并在平坦先验或 Student-t 先验下，从最大似然或贝叶斯视角对这些计数进行分析。接着，从每个合成数据集对应的后验分布中抽取一个样本，并计算该合成数据集对应的 p值。最后，我们将钠-葡萄糖协同转运蛋白2抑制剂治疗效应以及 $-\log_{10}$ 空间中 p/π 值的非参数核密度估计，与公式35和37给出的近似进行了比较。

图8所示的这些分析表明，当对数似然可以通过二次函数很好地近似时（CREDENCE/DAPA-CKD中的主要结局），公式35和37中引入的治疗效应预测分布及p/π值的闭式公式，是对这些量的实际预测密度（使用公式36精确模拟得到）的相当良好的近似。对于CREDENCE中的DKA结局，从5,000次重复模拟中获得的治疗效应经验预测分布的均值（标准误）为−7.15（30,000），因为其中许多数据集的观测值过少，无法进行分析，且估计值常常位于参数空间边界。因此，重复数据集中的p值围绕初始p值呈现相当分散的分布。公式37中的解析近似未能捕捉到接近1的p值所形成的高“尖峰”（图8，中排右列）。使用平坦先验或Student-t先验进行贝叶斯分析，都能够有效地对这些重复数据集的分析进行正则化，使这些重复π值更加集中地分布在初始p值附近。

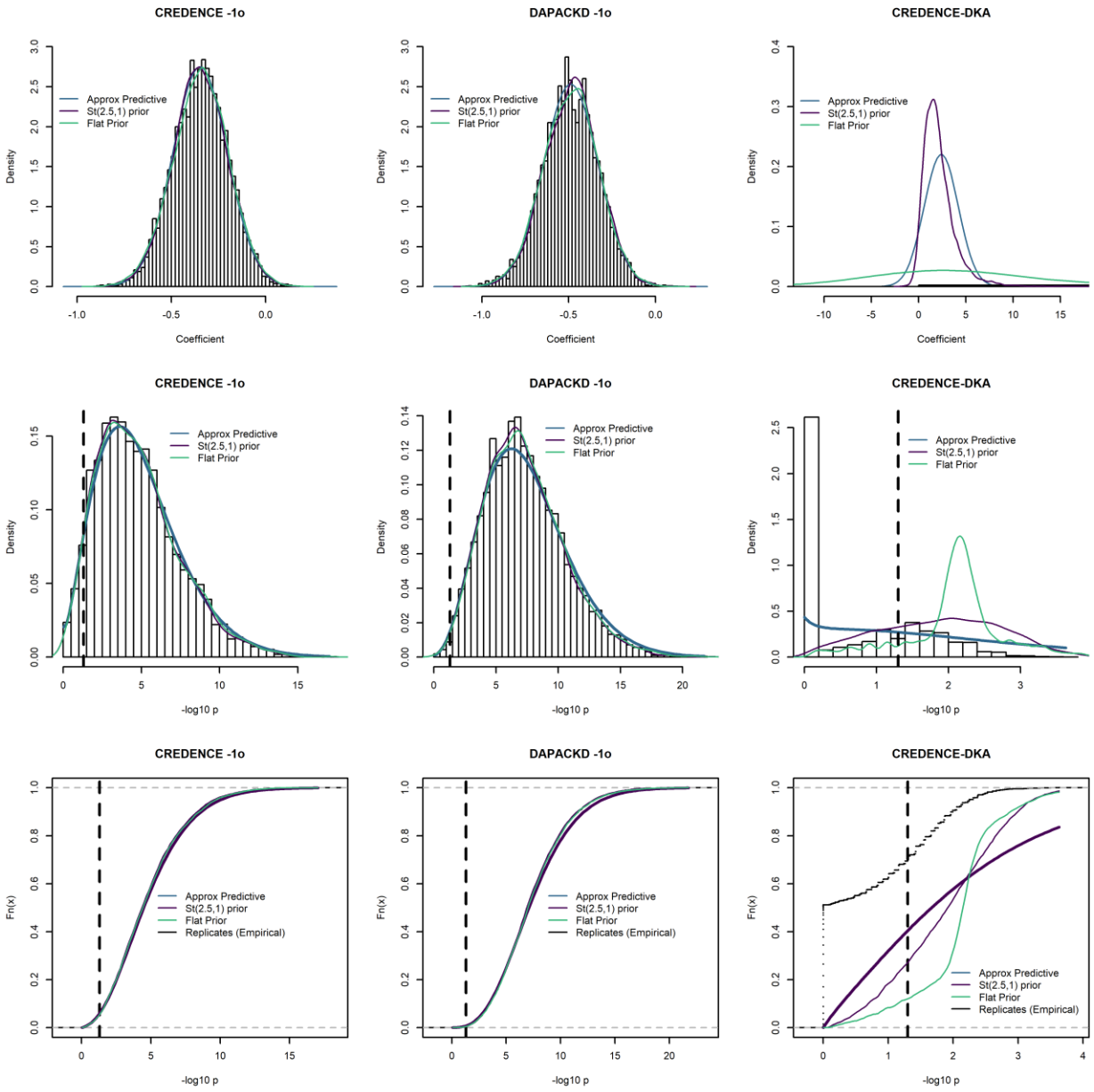


图*8 SGLT2*肾脏结局试验中治疗效应的复制分析。顶行：精确翻译和复制的试验中治疗效应预测分布的直方图、近似预测密度（公式*35*）用两种不同先验分布进行复制贝叶斯分析所得的非参数核密度估计。中行：复制试验中观测到的 *p* 值直方图、*p* 值和 *π* 值的近似预测密*37*），以及根据公式 *36* 所示方案，使用（基于 *NUTS* 采样器的）马尔可夫链蒙特卡罗进行复制贝叶斯分析所得的 *π* 值。对于贝叶斯分析研究中的后验样本拟合到有限高斯混合模型，并使用该模型的参数计算尾部概率。底行：基于公式 *37* 的 *p* 值和 *π* 值的累积分布函数。直虚线表示 *p=0.05* 的阈值。

7. 旨在重复验证的贝叶斯／似然主义综合分析

统计显著性检验的使用常常会被反复指认为科学事业中真实或虚构失败的罪魁祸首，例如每隔10至15年就会出现一次这样的指责；这些指责针对的是使用者使用统计检验的方式（Carver 1978；Chow 1988；Harris 1997；Nickerson 2000；Shao 和 Chow 2002；Wasserstein 等人 2019 ；Gibson 2021 ）。在最新一轮争论中，人们认为，将 p 值按 “0.05” 这一水平设定阈值，是科学领域重复验证危机的一个特别重要的成因（Siegfried 2010；Trafimow 和 Marks 2015；Gibson 2021）；随后，有人建议将显著性阈值降低至“<0.005”（Benjamin 等人 2018）——这距离 Edgeworth 使用这一数值已过去130多年（Kennedy-Shaffer 2019）；最终，美国统计学会针对统计推断中p值的误用发表了43篇文章，并在这些文章的回应中提议禁用“统计显著”一词（Wasserstein 等人 2019）。这一提议至今尚未实施；考虑到它在医学社论（Harrington 等人 2019 ）和临床试验期刊（ Cook 等人 2019 ）中遭到的反对，它恐怕也不会被采纳，这或许是因为“复杂的数学程序会给结论披上科学客观性的外衣”（Carver 1978）。在这场争论中，p值的拥护者并未保持沉默，他们指出，真正的问题在于统计显著性检验遭到误用，而非检验本身。这类误用源于“教科书和教程未能正确说明p值的推断含义”（Greenland 2019），本质上属于“人为因素问题”，可以通过“开发更好的循证教育和以用户为中心的统计软件”来应对（Lakens 2021）。简短（Senn 2001；Greenland 2019；Wasserstein 等人 2019）和长得多的禁忌清单（ Nickerson 2000；Greenland等人，2016年）也指出了使用者常犯错误的两个主要方面：习惯性地将p值误解为假设概率或错误概率；以及关注p值作为随机变量时的重复抽样性质，而不是某次分析中观测到的单一数值。或许“关键问题在于，这些概念没有任何一种解释能够

既简单直观，又正确无误 且万无一失”(Greenland等人，2016年) ，而且“高认知负荷导致了捷径式定义和解释的泛滥”。然而，p值相关问题以可预见的频率每隔10至15年就会反复出现，这一事实本身就表明，作为推断工具的p值与人类思维格格不入，对大多数科学家而言，它或许确实是一种相当不自然的推断方式。尽管如此，利用p值仍取得了大量科学研究成果，因此，尽管p值遭到了灾难性的滥用，其数值计算在科学推断中显然仍有价值。数值计算的效用与相关概念遭到系统性滥用之间的明显矛盾，亟待解释。我们认为，只要摒弃频率学派p值这一构造，并对广义线性模型（GLM）中计算并报告为p的数值赋予截然不同的贝叶斯解释，就能找到答案。我们将表明，在同一组数学假设下，通过GLM分析数据的贝叶斯学派和非贝叶斯学派，得到的后验分布要么完全相同，要么渐近等价。接着，我们将阐明计算后验尾部概率的理由，并建立一个对其进行阈值化的决策分析框架（π值）；对于任何给定的贝叶斯分析，这些数值都将与非贝叶斯p值完全相同。最后，我们将提出一种基于预测概率的论证来讨论可重复性。与p和π值的重复概率相比，这种论证不仅适用范围更广，还揭示了依赖后者衡量研究结果可重复性的严重局限。按照这一领域的惯例，我们将为统计学使用者列出一系列相互关联的“应做之事”和“禁忌”，说明贝叶斯π值与p值以往用法之间的关系，并强调将报告的p值重新表述为π值所需满足的技术要求。最后，我们将讨论改进统计分析报告的机会，以及在出版物和科学发现中对可重复性进行统计探索的机会。

1. **应使用π值来量化方向性（单变量）效应。**贝叶斯方法可以通过计算感兴趣参数符号的后验期望来量化方向效应，从而有效地将后者的取值范围映射到区间$[-1,1]$。π值越小，方向性效应的幅度越大。医学从业者会很快认识到，探究方向效应并据此采取行动，符合"施以帮助，或至少不造成伤害"这一伦理原则，即医学伦理学中"首先，不伤害"原则的原始表述（Smith 2005）。在药物开发语境下，监管机构在批准新疗法之前，可能只需知道其疗效优于对照治疗。因此，在医学和临床试验期刊中继续报告 p值（读者应可自由将其解读为π）仍应继续。

2. **务必使用 π-值对方向效应作出序数性主张。** 正如先前的评论者反复指出的那样，这一直是p值在应用研究中的一个重要应用领域（Chow 1988；Harris 1997；Nickerson 2000；Lakens 2021）。要理解原因，请注意，尽管人们会检验是否存在效应，但从来没有人会脱离效应的观测到的方向，孤立地解读统计显著性检验。例如，在一项优效性对照组 临床试验中报告显著的p值时，没有评论者会得出结论说，受检干预措施所得的数据只是与对照组所得数据不同（在统计意义上）。人们通常会依据观测到的p值评论观测到的效应方向，而科学出版物和交流中也一贯如此。然而，零假设检验并不正式允许作出此类方向性声明，只有方向性假设检验才允许。若要使科学家的报告与他们实际采用的统计程序相一致，就需要通过应用三值逻辑，对p值作出截然不同的重新诠释（Harris 1997）。在这种方法中，观测到的p值与关于参数幅度的陈述相联系

感兴趣的，即小于、难以判断（Hunter 1997），或大于零假设中指定的值。这种方案早在20世纪60年代初就已提出（Kaiser 1960），作为内曼-皮尔逊方法的扩展（顺便指出，贝叶斯文献在对方向性错误进行形式化分析方面相对较晚才出现，这类错误被称为“S”错误 （Gelman and Tuerlinckx 2000；Gelman and Carlin 2014；Lu 等，2019））。在非贝叶斯三值逻辑／方向推断框架中，可能出现三种不同的错误：声称存在差异而实际上并不存在（α类错误）、存在差异却未能发现（β类错误），或声称差异的方向错误（γ类错误）。采用这种方法会产生分尾显著性检验（Braver 1975；Harris 1997），在数学上将两个可能的单侧检验结合起来，用以检验方向性假设。虽然采用三值逻辑可以使解读与统计程序重新一致，但我们认为，我们提出的、涉及“重新校准”的分析师的框架，为科学家使用p值（适当解读为π-值）来声称存在方向性效应的行为提供了贝叶斯决策理论依据：分析师向客户提供的是对泛函的总结，即“信息片段”，这些总结基于后验分布以及一项直接针对客户决策问题的可行动性陈述（Hildreth 1963）。如果使用符号泛函，那么根据公式 28，传达的信息片段就是一个符号（效应方向），以及附加在该符号上的幅度（非贝叶斯分析中的检验统计量绝对值，或p值）。重要的是，无论π-值是否小于（“低于阈值”）决策程序的阈值，都可以作出此类主张。

3. **不要在涉及 π值的决策问题中使用自动阈值。** 频率学派假设检验中显著性水平的典型取值（例如 0.05 或 0.01），是手工计算尚属常态的时代的产物；当时出版成本限制了可复制并传播给科研工作者的表格篇幅（Stigler 2008；Kennedy-

Shaffer 2019)。选择 0.05，大致对应于估计效应与随机高斯噪声的均值相差两个标准差：乘以二很容易手工完成，无需计算机。Fisher 和 Neyman 都不认为阈值应当固定且普遍适用，而是认为它应视具体应用和领域而定。用后者的话来说（Neyman 1957）：“反对这种机械做法的理由很充分。事实上，显著性水平似乎应根据各种具体情形来确定，而这些情形因问题而异 ”。这些针对机械做法的警告，经过相应调整后，同样适用于 π值。值得考虑的是临界 p（频率学派）与 π （贝叶斯）阈值之间的关系。根据 Neyman（Neyman 1957）：“一方面，要考虑当所检验的假设实际上为真却被拒绝时，犯下错误（所谓第一类错误）的后果。根据问题的不同，这些后果可能十分严重，在这种情况下，就应采用更严格的显著性水平，或许是 0.001。确定显著性水平时还应考虑另一种情形：当所检验的假设实际上为假，且某个特定的备择假设为真时，却未能拒绝该假设（第二类错误）的后果。此外，还必须考虑这种错误发生的概率（或者说检验功效）。”因此，设定 p 的显著性阈值（即零假设为真时第一类错误的概率），似乎需要在判断真实状态所带来的收益与判断错误所造成的损失之间权衡。另一方面，π值的阈值则植根于更丰富的决策理论框架；该框架除了考虑决策本身的收益与损失，还会考虑推迟采取行动所产生的机会成本。本文写作之际，世界正经历 COVID-19 大流行，因此必须考虑统计决策中的收益、损失和机会成本，例如儿科疫苗试验的规模，或在试验中部署那些似乎能产生理想临床效果、但 p 值并非特别小的疗法

从未如此重要。我们提出公式 31，作为一种定量工具，以此开启关于如何针对具体问题和应用领域设定统计决策阈值的讨论。虽然 p值 的阈值也应有所变化，但我们认为，该理论中隐含的半定量权衡对如何将这些阈值的数值与下游应用联系起来，指导作用较弱。

4. **应使用低于阈值的 π值，视同存在方向性效应。**我们采用决策分析框架的唯一理由，正如 Neyman 所言，是为了“承担经过计算的风险，采取意志行动，以便未来（或许直到开展新的实验为止）以特定方式行事，并遵循实验的结局”（Neyman 1957）。第 4 节的决策框架旨在：当得到低于阈值的 π 值时，按照 Neyman 的思路作出可行动性陈述，并将新数据到达之日定为该陈述的失效日期。

5. **不要将超阈值的 π 值解读为不存在方向性效应的证据。** 表 4 所描绘的决策分析框架不包含“效应为零”这一行动/陈述，仅指出根据现有数据，该效应尚不可采取行动。有一种“恶意的零假设显著性检验邪教”（Greenland 2019），其成员“对研究的主要目标应是检验零假设这一观念，形成了强烈（且不健康）的关注”（Greenland等人，2016年）；所提出的框架可以避免陷入这种邪教的陷阱。事实上，我们的提议可视为回归统计显著性检验的历史基础（Edgeworth 1885）。Edgeworth 首次对显著性检验作出了正式的数学描述：在该检验中，观测到的 p 值用于确定某一特定证据总体的强度，以判断是否应认真对待某个问题（Stigler 2008; Kennedy-Shaffer 2019; Gibson 2021），也就是是否应据此采取行动。按照 Edgeworth 的说法，“或许值得采用一种方法，它不仅统一，而且可以说是不带个人色彩的，

而且还能够区分合理概率与近乎数学确定性之间的证据程度。” 在 Edgeworth 分析的例子中，近乎数学确定性相当于 0.005 这一现代阈值；然而，0.12 甚至 0.20 的数值则提供“一些证据，但其支持存在真实差异的力度远弱于我们在先前例子中所考察的证据”。在我们的决策框架中，π值未低于阈值，只会推迟作出“按效应存在来行动”的决定。不过，是否推迟进一步决策，取决于可采取行动的阈值。改变阈值只会改变在进一步考量科学发现之前所需的证据强度，但不一定会排除效应存在的可能性。获得超阈值的 π值时，可以利用效应量进一步探索后验分布，以概率方式回答方向性效应是否存在这一问题。

6. **请 将 π值及／或其与阈值的距离解读为方向性效应证据强度的衡量指标。** 这直接源于纯粹分析者与重新校准的分析者在完美信息的期望价值上的关系。简言之，π值越小，对于有关方向效应的序数性主张，任何额外信息的价值就越低。显然，这也与 Edgeworth 最初将 p 值用于区分证据强度的历史脉络相联系。

7. **如果数值等价的贝叶斯分析与非贝叶斯分析所需的数学假设均得到验证，务必将已发表的 p 值解读为 π 值。** 出于实际考虑，我们采用了广义线性模型，但数值等价所需的技术假设适用于更广泛的一类模型。这些假设可以方便地归纳为与似然和先验分布相关的两类。在似然方面，我们要求似然集中于远离参数空间边界的区域，对数似然在其最大值附近局部呈多元二次函数形式，并且参数向量的

有限维。 似然方面的关键假设是对数似然的“二次性”(Cam 和 Yang 2000；Geyer 2013)，它既使最大似然估计量趋于多元正态分布，又使我们能够通过拉普拉斯近似完成贝叶斯分析中的积分。在先验分布方面，主要假设是先验分布精确地或近似地为局部均匀先验；对于处理取值于实超立方体紧子集中的参数的分析师而言，这是一种从三个不同角度看都有理据的“客观”选择。贝叶斯主义者和非贝叶斯主义者都很喜欢把这类均匀先验视为 Laplace 时代的遗物，但我们认为，关于有限世界中推断、线性泛函诱导出均匀先验，以及不愿费心 eliciting 信息性先验的研究者群体所对应的极限形式的论据，足以证明在广义线性模型中采用这些先验是合理的。作出这些假设后，就可以应用拉普拉斯近似，推导出非贝叶斯方法与贝叶斯方法之间的数值等价。文献中已有一些相当精深的论述和分析，探讨指数族及具有二次对数似然的模型中后验分布的一致性，以及后验分布收敛至最大似然估计量的问题；Le Cam 所考察的模型包括局部渐近正态和局部渐近混合正态模型（Johnson 和 Ladalla 1979；DiCiccio 和 Stern 1993；Ghosal 等 1995；Ghosal 1997）。我们没有借助这些结果，而是出于教学目的进行了相关推导。据我们所知，迄今尚无人以一种能让读者清楚看出这一点的方式汇集这些材料：似然理论中一些最优雅的广义线性模型结果，实际上不过是在非信息性局部均匀先验下进行贝叶斯计算所得的低精度近似。认真研习渐近分析的读者还可以进一步验证，即使采用非均匀的弱信息先验（例如学生t分布），这种等价关系依然成立。相关数学工具基于优雅且

关于多元拉普拉斯近似中高阶项的简洁公式，例如（Bleistein and Handelsman 2010）中的隐式微分公式 8.3.50、Kirwin（Kirwin 2010）给出的定理1.2中的显式表示，以及针对指数族、统计学侧重点更强的论述（Ghosal et al. 1995; Ghosal 1997, 1999, 2000），尤其是（Ghosal 1997）中的定理2.1。唯一的要求是，对数似然比对数先验分布更尖锐，因此在渐近近似的高阶项中，前者的导数占主导地位。我们认为，对于在（局部）均匀先验分布或弱信息先验下贝叶斯分析与非贝叶斯分析之间的数值等价性（Senn 2001; Meng 2009; Lee et al. 2017），若不予承认，或仅仅口头上表示认可，就会严重损害那些希望成为贝叶斯主义者、但不习惯使用马尔可夫链蒙特卡罗的研究者的利益；同时，也会损害那些不得不采用多元正态近似的非贝叶斯主义者的利益，即使对数似然偏离二次函数形式也是如此。只要相关假设得到验证，前者就能够、也理应能够将最大似然计算“提升”到贝叶斯领域；而后者也不必因为启动马尔可夫链蒙特卡罗，并用其输出代替神奇 p* 公式而感到愧疚，因为该公式并未广泛应用于“现成的”统计软件中。对 SGLT2i 试验结局的分析提供了现实案例，足以为以非贝叶斯方式进行计算的贝叶斯主义者，以及反过来的情况，免除这些“罪过”。这些案例还表明，如果输入到对数似然呈二次函数形式的模型中的数据量足够大，那么即使样本量有限，也会达到“渐近无穷样本的状态”，因此先验分布不会影响推断。

8. **不要将 可重复性与相关科学发现的 π值混为一谈。** 我们在第5节中为广义线性模型推导出的边际预测分布，应当足以清楚地表明，p 和 π值仅具有较弱的可重复性，也就是说，它们并不会紧密聚集在初始研究所得的数值周围。实际上，尽管一种推断

意义应归于 基于单项研究计算出的尾部概率，再利用重复研究中的一系列设定阈值的 π 值来对科学发现的重复验证作出结论，但这种做法无异于愚蠢行径。类似的评论也适用于第5节中简要讨论的非贝叶斯预测框架下的 p 值。本节中的论证应当能够回应一个针对考察 p 值的可重复性概率这一做法所存在的主要隐含技术性质疑（Senn 2001、2002年；Greenland等人，2016年；Greenland 2019），即初始研究中获得的检验统计量在后续研究中也完全相同。公式 37 提供了一个相对简单的公式，用于在治疗效应进行精确转换且初始研究设计得到精确重复的情况下，计算 p 值和π值的 RPD。另一方面，公式36可用于探索更一般的设置，其中既涉及不完全迁移（例如偏倚），也涉及资源不足的重复验证（例如减少重复研究的样本量）同时考虑违反“二次性”的情况。 此类模拟将始终表明，在存在偏倚和/或用于重复验证研究发现的资源受到限制时，p 值和 π值的 RPD 将呈现明显更大的离散程度。这些考量进一步说明，应当摒弃这样一种倾向：将重复研究中获得的显著 p 值与不显著 p 值相对立，以此衡量发现的可重复性。生物医学和心理学应用中的大多数研究，其统计效能仅勉强足以检测其假定的效应，因此 p 值和 π值会在显著性阈值左右大幅波动。因此，最好不要对这类表面上的不一致反应过度，因为在实践中它们十分常见（Senn 2002）。

9. **不要使用公开可得的序列π-值来衡量某一领域中的全部证据。**这一点与前一点有些关联，但并不相同。许多领域的从业者常常会将大量数据归纳为摘要，统计探究某项发现的研究中产生显著 π-值的比例，并用这一比例来推断其中的

发现的可重复性。此处可以随意举出的这类陈述包括：“药物 X 有效，因为所有研究都发现结局 Y 有统计学显著降低”，或者“因素 W 的效应相当微弱，因为该领域只有一半的研究记录了统计学显著效应”。这种方法实际上将报告的 π值视为可用于以无偏方式进行元推断的量。若后一个假设得到验证，那么原则上可以使用 π值的绝对值，并将其拟合到对数正态分布，以便依照公式 37 尝试合并 π值，正如在费舍尔方法中所做的那样，即用于合并概率检验（Mosteller 和 Fisher，1948），或荟萃分析中使用的 Stouffer z分数法（Whitlock，2005年）。然而，关于某个主题发表的 p值很可能存在严重偏倚，这是因为“文件-抽屉”问题：如果实验 1 的结果显著，我们便在实验 2 中重复实验，若结果再次显著，一切就都没问题；但若不显著，我们就可能遇到不可重复性的情况，此时我们会通过诉诸 p值的分布来进行统计处理。但现实中会发生的是，如果实验 1 的结果不显著，该实验就会被塞进抽屉，永远不会发表。如果有多个团队在某个领域展开竞争，那么只有获得低于阈值的 π值的团队，才可能在转向更有前景的研究方向之前发表结果。此外，获得额外资源以重复验证高于阈值的 π值的可能性很低，而人们希望用这种荟萃分析方法利用的大量 π值可能永远不会产生，发表记录中只会保留低于阈值的值。这会导致对科学发现可重复性的评估更加偏斜、过度乐观。因此，按照定义，（不）可重复性的方向性偏倚在于只纳入低于阈值的 π值，同时它也极其依赖路径，取决于某项发现所在领域的早期历史（谁在何时发表了什么）。用于探查选择效应可能性的正式荟萃分析技术，例如治疗的漏斗图（Sterne 等，2011年）

效应及其精度，在证据综合之前，更有可能为评估可重复性提供有用信息。

10。 **当使贝叶斯与非贝叶斯分析在数值上等价的数学假设得到验证时，不要指望似然比检验能挽救针对二分式存在性（“无效应”）主张的检验。** 从方法论证伪主义的角度来看，针对效应是否存在的二分式统计检验似乎是必要的（有关这一观点的概述与辩护，见 Tunç 等，2021）。这一论点概括了关键实验（“决定性实验”）的思想：实验能够明确判定某一理论是否优于所有其他可能的假设。遗憾的是，如果要依据经过阈值化的 p 值来判定关键实验的结局，这一论点便站不住脚。在这种情况下，相应的统计量会与自由度等于所检验假设维数的 χ^2 进行比较，即似然比。若所检验的假设认为治疗效应向量的一个或多个分量为零，那么用这种检验来判定是否无效应，似乎是一个无可争议的立场。然而，完全置身于第 4 节框架内（该框架认为治疗效应始终存在）的人，只需用另一种离散化函数替换符号函数，就能得到完全相同的似然比检验值；该函数会将后验分布中与所检验假设（$\boldsymbol{\beta} = \boldsymbol{\beta_0}$）的马氏距离平方小于 $\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\right)^T \boldsymbol{\Sigma}^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\right)$ 的所有点映射为零，其余点映射为一。在平方损失函数下，该函数的后验期望等于多元正态函数在椭球区域上的积分的补数，而这恰好等于自由度适当的 χ^2 分布的尾部概率（p值）。纯分析师／客户和重新校准的分析师所面对的决策问题，也可以相应更新，以反映所引入的离散化映射。通俗地说，这种映射将参数的连续值离散化为零（“</output>

连续取值比观测到的值更接近假设的取值”）和一（“比观测到的值离点假设更远”），并将决策问题转化为对这一派生变量幅度的陈述。这些考量清楚地表明，一个人对效应是否存在的检验，在另一个人看来，就是对效应离散化幅度的陈述。除非离散化所用的马氏距离为无穷大（此时多元正态后验分布将与多元狄拉克δ函数重合），否则，依据具有超阈值 p 值的似然比检验所作的无效应陈述，其解释便值得商榷。在这种情况下，更准确的说法是：分析未能在用于分析数据的实验／分析的分辨能力（离散化）范围内检测到信号。

那么，在报告科学研究时，如何结合可重复性来实际运用这些注意事项？一个直接的解决方案应包含以下要素：

1. **在论文或报告的结果部分应采取似然主义者的立场，但在讨论部分则应采取贝叶斯立场。**科学论文实际上是向外部受众传达数据和推断的一种方式。贝叶斯分析与频率学派分析在数值上等价，这意味着，对于非贝叶斯广义线性模型分析所构建的任何汇总表，只要验证了二次性和远离边界的似然等相关假设，那么对同一数据进行贝叶斯分析时所构建的汇总表也会与之完全相同。从似然的角度呈现结果，却从贝叶斯的角度讨论相同结果，可以让使用者摆脱各种在数学上无关、因而不合理的解读禁令。例如，完全在贝叶斯框架内进行分析时，p 值（被解释为 π 值）并无争议，但后验汇总相对乏味，无法.....

治疗效应 基于粗略离散化，可以通过应用以下注意事项进行讨论。

2. **讨论当前发现时，应采用正式荟萃分析，将其置于既有研究发现的背景下，而不是比较 p 值。**鉴于 p 值在数值上的可重复性较差，将 p 值（以及 π 值！）与先前观测到的 p 值相对于显著性阈值进行比较的倾向，无非是“助长了二分癖这种强烈的认知偏差：即执意用二分法（‘非黑即白的思维’）取代数量，即使这种二分化对于推断并无必要，甚至会造成误导”，而且“会使使用者看不见实际的数据模式，从而令结论失效，有时甚至变得荒谬”（Greenland，2017年）。简单的荟萃分析或许能更有效地呈现当前研究为既有知识增添了什么。就此而言，在结果部分以似然主义者的身份行事，在讨论部分的前半部分以客观贝叶斯主义者的身份行事，并在讨论部分的中段以荟萃分析贝叶斯主义者的身份行事，可以清楚地区分以下任务：说明研究发现了什么、能够在脱离既有知识的情况下得出哪些推断，以及分析为全部证据增添了什么。

3. **运用预测框架，预先回应对本研究关键发现能否重复验证的疑虑。**通过公式35的近似公式，可以相当轻松地实现精确重复和翻译。这一论点仅适用于这种相当有限的情形；除非是在质量控制和审计机制都相当严格的高度规范化科学活动中，否则很难为这一计算提供合理依据。此类研究包括高能物理以及许多（但并非全部）药物和器械注册试验。在这些领域，应用公式35会得出诸如五西格玛规则这样的简单经验法则；如果 p 值足够小，甚至可以依据单项试验的结果批准药物。不幸的是，大多数研究并没有充足的资源来满足这些

要求，我们认为，相较于假设精确重复和翻译的公式，采用公式 36 的蒙特卡罗方案，考察不完全迁移和重复验证的可重复性，更值得提倡。讨论部分后段在讨论局限性时，是纳入此类分析的合适位置。翻译（“可推广性”）、数值近似（“二次性”）的局限性和重复验证（“小样本”）的局限性，可通过模拟框架进行相当 严格的探究，并呈现在稿件中。

4. **在方法部分和补充结果中，应包括对对数似然二次性的验证。**这是一项在实践中很少得到验证的数值假设，尽管它隐含于非贝叶斯广义线性模型计算中，并且对于将后者解释为近似贝叶斯计算是必需的。二次性可用于将针对当前推断似然任务拥有足够数据这一概念操作化（Geyer 2013），而且我们已经看到，对这一假设的验证可以使先验分布的选择与后验推断无关。基于最大值附近对数似然曲率的对数似然二次行为检验，最初是针对非线性最小二乘提出的（Bates and Watts 1980），作为对 Beale 的非线性度量（Beale 1960）以及 Box 提出的非线性回归偏倚表达式（Box 1971）的扩展。这些度量适用于广义线性模型（Kass and Slate 1994），并且可以从迭代加权最小二乘输出中计算出来，以补充对似然的直接可视化。如果无法验证二次性，通常将使似然主义结果失效，除非转而采用基于 p公式的方法。由于后者的数值实现并不广泛可用，而采用均匀先验的渐近等价马尔可夫链蒙特卡罗方法是可用的，因此转向贝叶斯数值方法可能是非贝叶斯主义者以数值上正确的方式开展分析的唯一途径。然而，一旦转而采用贝叶斯拟合程序，那么它将be

建议尝试多个不同的先验分布，以了解推断是由先验分布主导还是由数据主导（Kass 和 Wasserman，1996年）。SGLT2i 试验中罕见副作用的分析提供了“如何进行这类分析”的示例：CREDENCE 展示了成功的应用，而 DAPACKD 则提供了证据，表明有时甚至不应进行此类分析，直到其他研究获得更多数据。

总之，我们确立了广义线性模型的贝叶斯分析与非贝叶斯分析之间的数值等价关系，并提出了一些建议，说明如何利用这一等价关系改进分析结果的报告方式，同时在讨论和推断研究发现的可重复性时，摒弃不可靠的认知做法（例如 p值二分癖）。我们希望，这本简短专著所呈现的内容能为关于 p值作用的讨论作出实质性贡献，并为全面采用贝叶斯视角评估基于广义线性模型的证据指明方向。

附录： 在实数集中编程进行可重复性分析

在本附录中，我们将提供 R 分析软件代码，使用 SGLT2i 临床试验数据（第 6 节）对第节中可重复性的分析进行说明，并展示相关程序片段。假定读者熟悉 R编程语言。如果有意使用 MCMC 拟合广义线性模型，系统中还应安装可正常运行的 STAN 程序；它是一种功能强大的工具，可通过哈密顿蒙特卡罗和 NUTS采样器拟合贝叶斯模型。

A1. 可重复性计算

在实数集 *R* 中设置环境

我们首先加载拟合采用不同先验分布的模型并对结果进行后处理所需的库。

结果。这些库应在首次使用前安装：

```
库(arm)          ## 用于期望最大化贝叶斯拟合库(rstanarm)     ## 用于使用NUTS采样器（
```

输入并预处理这些分析所需的数据

现在，我们将表5中的数据输入 R，并选取 CREDENCE 中的主要结局作为示例（同一代码也可用于分析另外三个结局，但为节省篇幅，此处仅讨论 CREDENCE 中的主要结局）。在下面的数据框中，SGLT2i 试验中活性治疗（卡格列净或达格列净）组和安慰剂（PBO）组的事件数，以及以人年计的总随访时长，分别由 PYTrt 和 PYPBO 给出。

```
## 提供数据数据<-data.frame(结局=c("复合心肾结局", "糖尿病酮症酸中毒","复合心肾结
```

研究=c("CREDENCE","CREDENCE","DAPA-CKD","DAPA-CKD"), NGSLT2=

在 CREDENCE 中选择主要（复合心肾）结局 ## 用于说明 datExample<-subset(data, Outcome=="

使用贝叶斯方法和似然主义方法拟合广义线性模型

我们使用泊松广义线性模型分析钠-葡萄糖协同转运蛋白2抑制剂中的结局（事件数，NEvTrt 和 NEvPBO）

进行试验时，使用人年（PYTrt、PYPBO）作为偏置项。变量 Treat 给出的是 the

治疗分配（处理 = 1，临床试验中分配至SGLT2抑制剂组，处理 = 0 在此
安慰剂（PBO arm）。我们通过以下命令拟合泊松广义线性模型，并从似然主义拟合中提取 z 值：

通过最大似然拟合广义线性模型并提取用于计算 p值的 z 值 ## finit<-glm(计数~处理+偏置项(log(随

我们通过以下命令使用期望最大化方法拟合学生t分布先验（函数 *bayesglm* 是

可从软件包*arm*中获取）：

```
## 期望最大化拟合
finit.Bayes.EM<-bayesglm(计数~处理+偏置项(对数(随访时间)),数据=xinit,
                         family="泊松", prior.df=2.5,scale=1)
```

我们还可以使用 NUTS采样器，通过马尔可夫链蒙特卡罗拟合泊松广义线性模型。平坦先验分布和 Student-t 先验需要分别拟合。对于每种先验分布，我们拟合 4 条独立的马尔可夫链，并获得 20,000 个样本。在评估可重复性时，我们会从链中随机抽取子样本；大量迭代可确保估计值的蒙特卡罗误差保持较低，从而能够将系数比较到三位或四位有效数字。

使用平坦先验的 NUTS 拟合 - 设置种子以获得可复现的结果 finit.Bayes.flat<-stan_广义线性模型(计数~处理, 偏

student_t(df = 7, 0, 2.5), chains = 4, seed = 先验分布=截距 student_t(df = 7, 0, 2.5), iter =

要通过各方法获得结果，只需在 R 中调用相应的摘要方法：

摘要(finit) ## 最大似然结果

调用：
广义线性模型(公式 = 计数 ~ 处理 + 偏置项(对数(FUT)), 族 = "泊松",
    数据 = xinit)

偏差残差：
[1] 0 0

系数：
估计值 标准误 z 值 Pr(>|z|)
(截距) 4.11415 0.05423 75.861 < 2e-16 ***
处理 -0.34831 0.08380 -4.156 3.23e-05 ***

显著性代码：0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(泊松族的离散参数取值为 1)

零偏差：1.7517e+01 ，自由度为 1
残差偏差：2.7756e-14，自由度为 0
赤池信息准则：

费舍尔评分迭代次数：2

摘要(finit.Bayes.EM) ## 期望最大化贝叶斯结果

调用：

bayesglm(公式 = 计数 ~ 处理 + 偏置项(对数(随访时间)), family = "泊松", 数据 = xinit, prior.df = 2.

偏差残差:
[1] 0 0

系数:
估计值 标准误 z 值 Pr(>|z|)
截距 4.11382 0.05422 75.868 < 2e-16 ***
处理 -0.34780 0.08374 -4.153 3.28e-05 ***

显著性代码: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(泊松族的离散参数取为 1)

零偏差: 1.7517e+01 , 自由度为 1; 残差偏差: 4.3978e-05, 自由度

费舍尔评分迭代次数: 4

摘要(finit.Bayes.flat,digits=4) ## NUTS - 平坦先验结果

模型信息:

function: stan_glm
族: 泊松 [对数]
公式: 计数 ~ 处理
算法: 抽样
先验分布: 参见 help('prior_summary')
样本: 40000 (后验样本量)
观测值: 2
预测变量: 2

估计值:
均值 标准差 2.5% 25% 50% 75% 97.5% (截距) 4.1125 0.0541 4.0051 4.0762 4.1132 4.1496

诊断:
蒙特卡洛标准误 Rhat n_eff
截距 0.0003 0.9999 28495
处理 0.0005 1.0001 26253
mean_后验预测分布 0.0962 1.0000 31856
对数后验 0.0077 1.0001 16609

对于每个参数, mcse 表示蒙特卡洛标准误, n¯eff 是有效样本量的粗略衡量指标, Rhat 是分裂链上的

摘要 (finit.Bayes,digits=4) ## NUTS - 学生t分布先验分布结果

模型信息:

函数: stan glm_

分布族：泊松 [ ] 对数公式：计数 处理 算法：抽样 先验分布：参见 help('prior summary') 样本：40000（后验样本量）观测值：2 预测变量：2

估计值：

	均值	标准差	2.5%	25%	50%	75%	97.5%
（截距）	4.1097	0.0543	4.0023	4.0733	4.1103	4.1470	4.2135
处理	-0.3439	0.0840	-0.5092	-0.4001	-0.3442	-0.2872	-0.1797
平均后验预测分布	292.1201	17.1153	259.5000	280.5000	292.0000	303.5000	326.5000
对数后验	-13.9318	0.9958	-16.6100	-14.3169	-13.6291	-13.2240	-12.9579

诊断：

	蒙特卡洛标准误	Rhat	n 有效样本量
0001 2925 （截距）	0.0003	1.	
处理	0.0005	1.0000	24755
平均后验预测分布	0.0967	1.0000	31300
对数后验	0.0075	1.0002	17582

对于每个参数，mcse 是蒙特卡洛标准误，n_eff 是有效样本量的粗略度量

规模，而 Rhat 是分裂链上的潜在尺度缩减因子（收敛时 Rhat=1）。π值最好通过平滑 MCMC 的输出，而不是直接使用公式 28 中的样本来计算贝叶斯模型的 p 值。要理解原因，请注意，根据 MCMC 样本的经验分布估计的 π值下限为 2/(MCMC 中的样本数)。因此，要将 π值计算到 3、4、5 位有效数字，分别需要>1、1,0000、100,000 个样本；而通过平滑程序得出的估计值可能更精确。可以将低维高斯混合拟合到链的输出，再利用混合各成分观测到的混合比例和 z值，通过代数方法计算 p 值，从而得到基于平滑的 p 值估计值。以下代码将对一个 MCMC 样本拟合单变量高斯混合并执行相关计算；我们对每个参数分别拟合，因为在高维情况下，由于维数灾难，MCMC 输出中的样本可能不足以拟合高维混合并捕捉尾部行为。从形式上说，我们利用这样一个事实：针对多个参数的联合马尔可夫链蒙特卡罗模拟，也会从边缘分布中抽取样本

每个参数的分布。 因此，我们可以大幅简化 π值的计算，方法是

以边缘方式，即单变量方式进行。

使用高斯混合从 MCMC 后验分布近似尾部概率 pval.MCLUST<-function(mcmc.df) { kde.df<-

我们首先将　　MCMC样本转换为矩阵，然后对每一列应用上述函数

：

```
## 平坦先验分布  bayes.flat<-as.matrix(finit.Bayes.flat)          ## 将样本放入矩
```

```
Student-t先验
贝叶斯<-as.matrix(finit.Bayes)                    ## 将样本放入矩阵
p.bayes<-应用(bayes,2,pval.MCLUST)              ## 计算p值
```

与直接从经验分布计算 π 值相比：

经验p值 ## 平坦先验 p.bayes.flat.emp<-apply(bayes.flat,2,function(x) 2*min(均值(x>=0),均值(x<0)))

由于 MCMC 链中只有 80,000 个样本（4 条链，每条链包含 20,000 个样本），因此采用经验方法能够得到的最小非零 π-值为 2/80,000 = 5 10^{-5}。然而，截距的实际 π-值小于这一数值，因此经验计算结果直接下溢为零。另一方面，π-值大于计算精度的下限，因此经验 π-值

```
rbind(p.bayes.flat.emp, p.bayes.flat)
```
截距 处理

```
p.bayes.flat.emp 0.000000e+00 0.001050000
p.bayes.flat     1.751247e-05 0.001895736
```

```
rbind(p.bayes.emp, p.bayes)
```
截距 处理
p.bayes.emp 0.000000e+00 0.00860000
p.bayes 2.295234e-06 0.01161849

要实现用于可重复性的边际预测框架（公式36），我们必须从平坦先验分布下得到的后验分布中进行模拟，并使用抽取的系数生成一个新数据集，再用它进行估计。初始样本后验分布中抽取的

模拟泊松计数 ## 注意，改变随访时间（FUT）相对于初始研究中 ## 观察到的值，可用于

现在拟合贝叶斯模型 ## NUTS - 平坦先验finit.Bayes.flat<-stan_glm(计数~处理, 偏置项

数据=dsim, family=泊松(link="对数"), 先验分布 = 学生t分布

(df =) iter) bayes<-as.matrix(f

在重复数据集中反复应用 NUTS采样器时，有可能出现这种情况：

对于一个或多个此类数据集，算法将无法收敛。我们使用了 R 的异常处理机制，

通过包装一个捕获此类情况的函数，跳过马尔可夫链蒙特卡罗未能收敛的数据集，若出现此类情况则.....

在真实模拟函数上的误差：

STAN 的异常处理封装

```
容错仿真<-函数(b0,b1)
{
  <escape>tryCatch({ret <- sim(b0,b1);}, 错误 = 函数(e) {ret <<- rep(NA,12)});</escape>
  ret

}
```

分析中会生成有限数量（例如 5000 个）的重复数据集

以评估可重复性。以下代码将通过对 MCMC 链进行子抽样来生成这些数据集

在平坦先验下分析原始数据集中的数据所得：

set.seed(12345) ## 设置随机种子以获得可重复的结果 Nsim<-5000 ## 要获取的蒙特卡罗样本数量

以下内容的代码 对π值概率密度（公式 37）及其在以 10 为底的对数尺度下的累积分布函数，给定在初始阶段获得的统计量值

数据集如下：

```
## 近似 p 值累积分布函数
## T 是初始样本中观察到的统计量，## log10p 是我们要计算累积分布函数
```

```
## 近似 p 值概率密度函数 pvalPDF<-function(log10p,T) { c<-qnorm((10^
```

A2. 将计算结果绘图

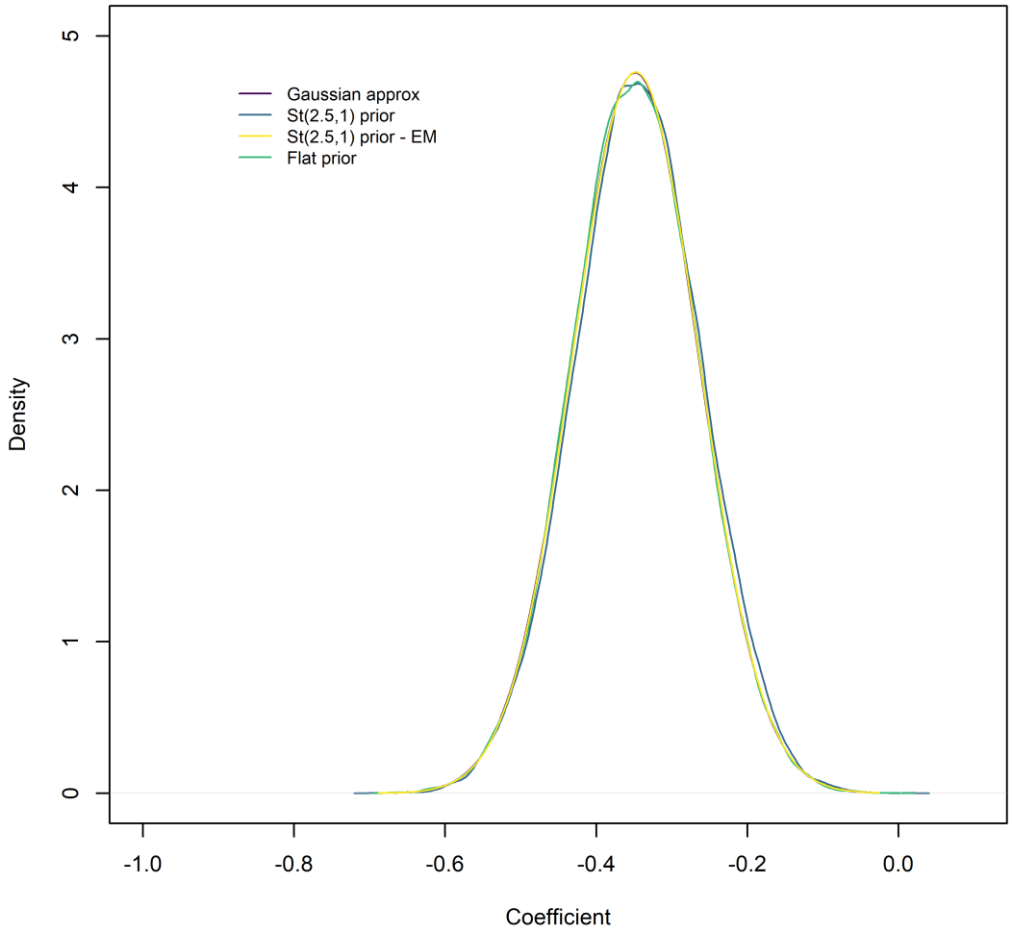
通过图形对比不同方法在初始数据集中得到的系数估计值（与最大似然方法一致的高斯近似、通过期望最大化拟合的 Student-t 先验分布，或采用两种不同先验分布的 NUTS），以检测其偏离正态性的情况。以下代码片段可以实现这一点：

```
col=viridis(4) ## 获取四种适合色盲的颜色调色板 main="CREDENCE -1o" ## 总体图表标签 x<-
```

Student-p 先验，期望最大化算法 ## 需要一些编程操作才能从对象中提取系数 em.coef<-

```
先验分布"),bty="n",xjust=1,lty=1,cex=0.8)
```

由此生成的图形如下：



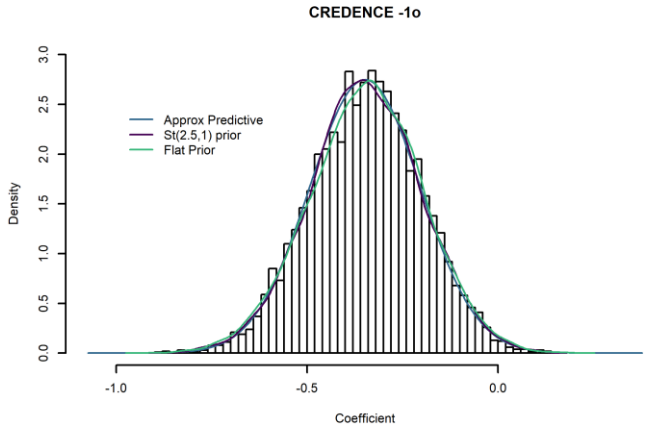
以下代码以图形方式汇总了通过最大似然对重复数据集进行的分析，以及在平坦先验分布和 Student-t 先验下的 MCMC 估计。

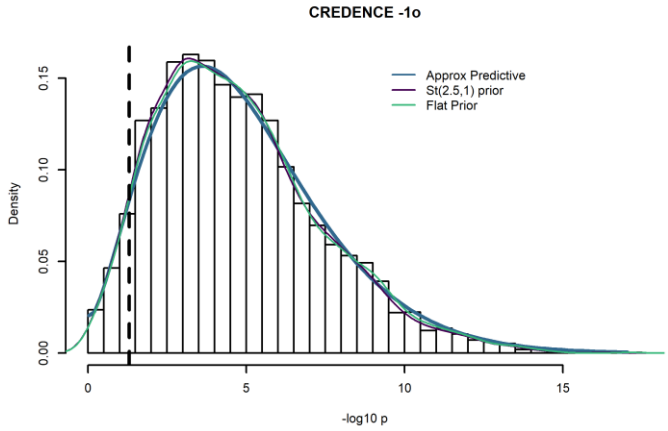
hist.algo<-"fd" ## 用于直方图评估的算法 par(mfcol=c(3,1)) x<-seq(quantile(Best[,"B1"],probs=0.

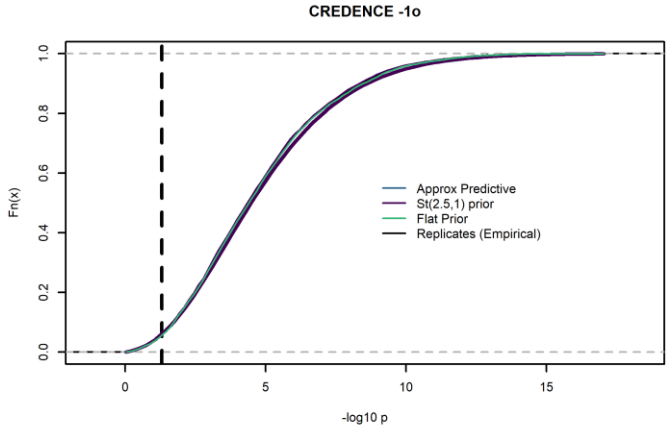
s

hist(Best[,"B1"],freq = F, breaks = hist.algo, add=F, main=main, xlab="系数", ylab="密度",

代码将生成三幅图，分别显示系数、重复数据的概率密度函数π及其累积分布函数。下方展示了 CREDENCE 研究中主要结局的输出：







参考文献

Argyropoulos C，Unruh ML（2015年）利用广义加性模型分析随机对照试验中的事件时间结局。PLoS ONE 10:e0123784。https://doi.org/10.1371/journal.pone.0123784

Balasubramanian V（1997年）概率分布空间上的统计推断、奥卡姆剃刀与统计力学。Neural Computation 9:349–368。https://doi.org/10.1162/neco.1997.9.2.349

Barndorff-Nielsen O（1983年）关于最大似然估计量分布的一个公式。Biometrika 70:343–365。https://doi.org/10.2307/2335549

Barndorff-Nielsen O（1980年）条件性解。Biometrika 67:293–310。https://doi.org/10.1093/biomet/67.2.293

Barndorff-Nielsen O, Cox DR（1979年）《Edgeworth 展开与鞍点近似及其统计应用》。《皇家统计学会期刊 B辑（方法论）》41：279–312

Bates DM, Watts DG（1980年）《非线性的相对曲率度量》。《皇家统计学会期刊 B辑（方法论）》42：1–25

Beale EML（1960年）《非线性估计中的置信域》。《皇家统计学会期刊 B辑（方法论）》22：41–88

Bedrick EJ, Christensen R, Johnson W（1996年）《广义线性模型先验分布的新视角》。《美国统计学会杂志》91：1450–1460。https://doi.org/10.1080/01621459.1996.10476713

Bender A, Groll A, Scheipl F（2018年）《广义加性模型在事件时间分析中的应用》。Statistical Modelling 1471082X17748083。https://doi.org/10.1177/1471082X17748083

Benjamin DJ, Berger JO, Johannesson M 等（2018年）重新定义统计显著性。Nat Hum Behav 2:6–10. https://doi.org/10.1038/s41562-017-0189-z

Benjamini Y, Hochberg Y（1995）控制假发现率：一种实用且有效的多重检验方法。皇家统计学会期刊 B辑（方法论）57:289–300

Berger J（2006）支持客观贝叶斯分析。贝叶斯分析 1:385–402. https://doi.org/10.1214/06-BA115

Berger JO（2003）Fisher、Jeffreys 和 Neyman 能否就检验达成一致？Statist Sci 18:1–32. https://doi.org/10.1214/ss/1056397485

Blæsild P, Jensen JL (1985) 再生指数模型的鞍点公式。《斯堪的纳维亚统计学杂志》12：193–202

Bleistein N, Handelsman RA（2010）《积分的渐近展开》，修订版。Dover 出版物，纽约。Boos DD, Stefanski LA（2011）《P值的精确性与可重复性》。《美国统计学家》65:213–221。https://doi.org/10.1198/tas.2011.10129 Bornkamp B（2012）《非线性建模的函数型均匀先验》。《生物统计学》68:893–901 Box GEP, Tiao GC（1973）《贝叶斯推断的本质》。载于：《统计分析中的贝叶斯推断》。第1–75 Box MJ（1971）《非线性估计中的偏倚》。《皇家统计学会期刊 B辑（方法论）》33:171–201 Braver SL（1975）《关于不等分配尾部：单尾与双尾检验的新视角》。《教育与心理测量》35:283–301。https://doi.org/10.1177/001316447503500206 Buse A（1982年）《似然比、Wald和拉格朗日乘数检验：说明性注释》。《美国统计学家》36:153–157。https://doi.org/10.2307/2683166 Cam LL, Yang GL（2000）《统计学中的渐近理论：若干基本概念》，第2版。施普林格出版社，纽约。Campbell L（1966）《作为分布范围度量的指数熵》。《概率论及相关领域》5:217–225 Carver R（1978）《反对统计显著性检验》。《哈佛教育评论》48:378–399。https://doi.org/10.17763/haer.48.3.t490261645281841 Chen B, Wang J, Zhao H, Principe JC（2016）《关于熵作为多变量变异性度量的见解》

.

《熵》18:196. https://doi.org/10.3390/e18050196 Chow SL（1988年）显著性检验还是效应量？《心理学公报》103:105–110. https://doi.org/10.1037/0033-2909.103.1.105 Colquhoun D（2017年）研究的可重复性及对p值的误读。《英国皇家学会开放科学》4:171085. https://doi.org/10.1098/rsos.171085 Colquhoun D（2014）对假发现率及对p值误读的调查。《英国皇家学会开放科学》1:140216. https://doi.org/10.1098/rsos.140216 Cook JA, Fergusson DA, Ford I 等（2019年）社论：临床研究中仍有显著性检验的一席之地

al

试验。《临床试验》16:223–224。https://doi.org/10.1177/1740774519846504 Cordeiro GM（2004）广义线性模型中最大似然估计的二阶协方差矩阵。《统计与概率快报》66:153–160。https://doi.org/10.1016/j.spl.2003.10.010

Cordeiro GM（1983）广义线性模型的改进的似然比统计量。《皇家统计学会学报 B辑（方法学）》45:404–413 Daniels HE（1954）统计学中的鞍点近似。《数学统计年刊》25:631–650。https://doi.org/10.1214/aoms/1177728652 De Capitani L、De Martini D（2016）某些非参数检验的可重复性概率估计与RP检验。《熵》18:142。https://doi.org/10.3390/e18040142 De Martini D（2008）统计假设检验的可重复性概率估计。《统计与概率快报》78:1056–1061。https://doi.org/10.1016/j.spl.2007.09.064

Denham R、Mengersen K（2007）回归模型中专家意见的地理辅助信息征询。《贝叶斯分析》2：99–135。https://doi.org/10.1214/07-BA205

DiCiccio TJ，Stern SE（1993）关于近似贝叶斯推断的巴特利特校正。Biometrika 80：731–740。https://doi.org/10.2307/2336865

Dobson AJ, Barnett AG（2018年）《广义线性模型导论》，第4版。Chapman and Hall/CRC，博卡拉顿

Durbin J（1980年）充分估计量的密度近似。Biometrika 67:311–333。https://doi.org/10.2307/2335475

Edgeworth FY（1885）《统计学方法》。《伦敦统计学会期刊》，181–217。

Edwards W, Lindman H, J. Savage L（1963）《心理学研究中的贝叶斯统计推断》。《心理学评论》70:193–242。https://doi.org/10.1037/h0044139

Efron B（1998年）21世纪的 R. A. Fisher（在1996年 R. A. Fisher讲座上发表的特邀论文）。Statist Sci 13:95–122. https://doi.org/10.1214/ss/1028905930

EfronB（1988年）逻辑回归、生存分析和Kaplan-Meier曲线。《美国统计学会杂志》83:414–425

Engle RF（1984）第13章：计量经济学中的沃尔德检验、似然比检验和拉格朗日乘数检验。载于：《计量经济学手册》。Elsevier，第775–826页。

Fulks W（1960年）《渐近分析 I：关于拉普拉斯方法的一则注记》。《美国数学月刊》67：880–882。https://doi.org/10.2307/2309456

Fulks W、Sather JO（1961）渐近分析。II. 重积分的拉普拉斯方法。《太平洋数学杂志》11：185–192

Galak J, LeBoeuf RA, Nelson LD, Simmons JP（2012年）纠正过去：未能重复验证 psi。《人格与社会心理学杂志》103:933–948。https://doi.org/10.1037/a0029709

Garthwaite PH, Al-Awadhi SA, Elfadaly FG, Jenkinson DJ（2013年）广义线性模型和分段线性模型的先验分布信息征询。《应用统计学杂志》40:59–75。https://doi.org/10.1080/02664763.2012.734794 Gelman A, Carlin J（2014年）超越功效计算：评估 S 型（符号）错误和 M型（幅度）错误。《心理科学展望》9:641–651。https://doi.org/10.1177/1745691614551642 Gelman A, Jakulin A, Pittau MG, Su Y-S（2008年）用于逻辑斯蒂

c

及其他回归模型。《应用统计年鉴》2：。https://doi.org/10.1214/08-AOAS191 Gelman A，Tuerlinckx F（2000）经典和贝叶斯单重及多重比较程序的 S 型错误率。《计算统计学》15：373–390。https://doi.org/10.1007/s001800000040 Geyer CJ（2013）无需大数定律、中心极限定理或趋于无穷大的样本量的极大似然渐近理论。载于：《数理统计学会文集》。数理统计学会

,

美国俄亥俄州比奇伍德，第 1–24 页 Ghosal S（1997）具有众多协变量的广义模型后验分布的正态近似。《统计学数学方法》6：332–348

Ghosal S（2000）参数数量趋于无穷时指数族后验分布的渐近正态性。《多元分析杂志》74：49–68。https://doi.org/10.1006/jmva.1999.1874

Ghosal S（1999）高维线性模型中后验分布的渐近正态性。《伯努利》5：315–331 Ghosal S、Ghosh JK、Ramamoorthi RV（1997）通过筛与覆盖数构造非信息先验。I.

n:

统计决策理论及其应用进展。Birkhauser Boston Inc.，美国，第 119–132 页。Ghosal S，Ghosh JK，Samanta T（1995）论后验分布的收敛。《统计年鉴》

cs

23:2145–2152 Ghosh JK（1994年）第8章：魔法公式、巴特利特校正与匹配概率。载于：《高阶渐近理论》。数理统计学会与美国统计学会，加利福尼亚州海伍德和弗吉尼亚州亚历山德里亚，第74–85页。Gibson EW（2021年）《p值在评判证据强度和现实重复验证预期中的作用》。《生物制药研究统计学》13：6–18。https://doi.org/10.1080/19466315.2020.1724560Goodman SN（1992年）《关于重复验证、p值与证据的评论》。《医学统计学》11：875–879。Goodrich B、Gabry J、Ali I、Brilleman S（2020年）《rstanarm：通过Stan进行贝叶斯应用回归建模》。

Greenland S（2019年）有效的 P 值的表现完全符合预期：对 P 值的一些误导性批评及其通过 S 值的解决方案。《美国统计学家》73：106–114.https://doi.org/10.1080/00031305.2018.1529625

Greenland S（2017年）特邀评论：方法学需要认知科学。《美国流行病学杂志》186：639–645。https://doi.org/10.1093/aje/kwx259

Greenland S、Senn SJ、Rothman KJ 等（2016年）统计检验、p 值、置信区间和检验效能：误读指南。《欧洲流行病学杂志》31：337–350。https://doi.org/10.1007/s10654-016-0149-3

Greenwald AG、Gonzalez R、Harris RJ、Guthrie D（1996年）效应量和 p 值：应报告什么，又应重复验证什么？《心理生理学》33：175–183

Grieve A（1988年）分析 LD50 实验的贝叶斯方法。《贝叶斯统计学》3：617–630

Harlow LL、Mulaik SA、Steiger JH（编）（2013年）《假如不存在显著性检验？》纽约：Psychology Press

Harrington D、D'Agostino RB、Gatsonis C 等（2019年）《期刊统计报告新指南》。N Engl J Med 381:285–286。https://doi.org/10.1056/NEJMe1906559

Harris R（1997年）通过三值逻辑改革显著性检验。载于：《如果没有显著性检验会怎样？》劳伦斯·埃尔鲍姆出版社，新泽西州马瓦，145–174页。

Hayakawa T（1975）局部备择假设下复合假设的似然比准则。Biometrika 62：451–460。https://doi.org/10.2307/2335387

Heerspink HJL、Stefánsson BV、Correa-Rotter R 等（2020年）达格列净治疗慢性肾脏病患者。N Engl J Med 383:1436–1446。https://doi.org/10.1056/NEJMoa2024816

Hildreth C（1963）贝叶斯统计学家与远程客户。Econometrica 31:422–438。https://doi.org/10.2307/1909982

Hinkley DV（1980年）似然作为近似枢轴分布。Biometrika 67:287–292。https://doi.org/10.1093/biomet/67.2.287

Hoffman MD、Gelman A（2014）无 U 形转向采样器：自适应设定哈密顿蒙特卡罗中的路径长度。Journal of Machine Learning Research 15:1593–1623

Hosack GR、Hayes KR、Barry SC（2017年）贝叶斯广义线性模型的先验 elicitation及其在风险控制方案评估中的应用。Reliability Engineering & System Safety 167:351– 361。https://doi.org/10.1016/j.ress.2017.06.011

Hunter JE（1997年）亟需：禁止显著性检验。Psychol Sci 8:3–7。https://doi.org/10.1111/j.1467-9280.1997.tb00534.x

Ibrahim JG、Laud PW（1991年）关于使用 Jeffreys 先验分布对广义线性模型进行贝叶斯分析。《美国统计学会杂志》86：981–986。https://doi.org/10.2307/2290514

Ioannidis JPA（2005年）为何大多数已发表的研究发现都是错误的。PLOS Medicine 2:e124。https://doi.org/10.1371/journal.pmed.0020124 Jeffreys H （ 1946 年）估计问题中先验概率的不变形式。《伦敦皇家学会数学与物理科学系列论文集》186：453–461 Johansen ME、Argyropoulos C（2020年）心血管结局、心力衰竭和肾病试验表明，现在是使用钠-葡萄糖协同转运蛋白2抑制剂的时候了。Clinical Cardiology 43：1376–1387。https://doi.org/10.1002/clc.23508

Johnson RA、Ladalla JN（1979年）从多参数指数族模型抽样时后验分布的大样本行为及相关结果。《Sankhyā：印度统计学杂志，B辑》（1960年—2002年）41：196–215

Jorgensen B（1997年）《离散模型理论》。CRC出版社。

Jorgensen B（1987年）指数色散模型。《皇家统计学会期刊 B辑（方法论）》49：127–162

Jørgensen B（1992年）指数离散模型及其扩展：综述。International Statistical Review / Revue Internationale de Statistique 60：5–20。https://doi.org/10.2307/1403498

Jørgensen B（1987年）小离散渐近理论。Brazilian Journal of Probability and Statistics 1：59–90

Kadane J, Wolfson LJ（1998年）信息征询的经验。《皇家统计学会学报：D辑（统计学家）》47：3–19。https://doi.org/10.1111/1467-9884.00113

Kadane JB、Dickey JM、Winkler RL 等（1980）。正态线性模型的意见交互式信息征询。《美国统计学会杂志》75：845–854。https://doi.org/10.1080/01621459.1980.10477562

Kaiser HF（1960年）方向性统计决策。《心理学评论》67：160–167。

Kass R, Tierney L, Kadane J（1990年）基于拉普拉斯方法的后验展开的有效性。载于：Geisser S、Hodges J、Press S、Zellner A（编）《统计学与计量经济学中的贝叶斯与似然方法》。North Holland

Kass RE，Slate EH（1994年）最大似然和后验非正态性的一些诊断。Ann Statist 22：. https://doi.org/10.1214/aos/1176325490

Kass RE，Wasserman L（1996年）依据形式规则选择先验分布。《美国统计学会杂志》91：1343–1370。https://doi.org/10.2307/2291752

Kennedy-Shaffer L（2019）从 $p < 0.05$ 之前到 $p < 0.05$ 之后：运用历史背景理解 p 值与显著性检验。《美国统计学家》73:82–90。https://doi.org/10.1080/00031305.2018.1537891 Kirwin WD（2010）拉普拉斯近似的高阶渐近。《渐近分析》70:231–248。https://doi.org/10.3233/ASY-2010-1016

Lakens D（2021）p 值的实用替代方案是正确使用 p 值。《心理科学展望》16:639–648。https://doi.org/10.1177/1745691620958012 Lambert D, Hall WJ（1982）p 值的渐近对数正态性。《统计年鉴》10:44–64。https://doi.org/10.1214/aos/1176345689 Lee Y, Nelder JA, Pawitan Y（2017）含随机效应的广义线性模型：通过 H 似然进行统一分析，第二版，第 2 版。Chapman and Hall/CRC，博卡拉顿，佛罗里达州 Lindsey JK（1997）《广义线性模型的应用》，修订版。Springer，纽约 Lu J, Qiu Y, Deng A（2019）关于假设检验中 S/M 类错误的说明。《英国数理统计心理学杂志》72:1–17。https://doi.org/10.1111/bmsp.12132 McCullagh P, Nelder JA（1989）《广义线性模型》，第二版。Chapman & Hall

Meehl PE（1978）《理论风险与表格中的星号：卡尔爵士、罗纳德爵士与软心理学的缓慢进展》。《咨询与临床心理学杂志》46：806–834

Meehl PE（1967）《心理学与物理学中的理论检验：一个方法论悖论》。Philosophy of Science 34:103–115 Meng X-L（2009）《解码 H-似然》。Statist Sci 24:280–293。https://doi.org/10.1214/09-STS277C Morey RD、Hoekstra R、Rouder JN 等（2016a）《对置信区间寄予信任的谬误》。Psychon Bull Rev 23:103–123。https://doi.org/10.3758/s13423-015-0947-8 Morey RD、Romeijn J-W、Rouder JN（2016b）《贝叶斯因子的哲学与统计证据的量化》。Journal of Mathematical Psychology 72:6–18。https://doi.org/10.1016/j.jmp.2015.11.001 Mosteller F、Fisher RA（1948）《问题与答案》。The American Statistician 2:30–31。https://doi.org/10.2307/2681650 Nelder JA、Pregibon D（1987）《扩展准似然函数》。Biometrika 74:221–232。https://doi.org/10.2307/2336136 Neyman J（1957）《“归纳行为”作为科学哲学的基本概念》。Revue de l’Institut International de Statistique / Review of the International Statistical Institute 25:7–22。https://doi.org/10.2307/1401671 Nickerson RS（2000）《零假设显著性检验：一场持久争议的回顾》。Psychological Methods 5:241–301。https://doi.org/10.1037/1082-989X.5.2.241 Pashler H、Wagenmakers E（2012）《心理科学可重复性专题栏编者导言》。Perspectives on Psychological Science：《心理科学展望》：美国心理科学协会期刊。https://doi.org/10.1177/1745691612465253

Pawitan Y（2013）《一切皆似然：使用似然进行统计建模与推断》。牛津 大学出版社，牛津，纽约 Peña D、Rodrí´guez J（2003）《多元散布与线性依赖的描述性度量》。Journa

l

《多元分析》85:361–374。https://doi.org/10.1016/S0047-259X(02)00061-1
Perkovic V、Jardine MJ、Neal B 等（2019年）卡格列净与2型糖尿病及肾病患者的肾脏结局。《新英格兰医学杂志》380:2295–2306。https://doi.org/10.1056/NEJMoa1811744

Racine A、Grieve AP、Fluhler H、Smith AFM（1986年）实践中的贝叶斯方法：制药行业的经验。《皇家统计学会杂志C辑》（应用统计）35:93–150。https://doi.org/10.2307/2347264

Reid N（1988年）鞍点方法与统计推断。Statist Sci 3:213–227。
https://doi.org/10.1214/ss/1177012906

Senn S（2001年）为p值喝彩两声？《流行病学与生物统计学杂志》6：193–204。
https://doi.org/10.1080/135952201753172953

Senn S（2002年）关于重复验证、p值与证据的评论，S.N.Goodman，《医学统计学》1992年；11：875-879。《医学统计学》21：2437–2444；作者回复2445-2447。
https://doi.org/10.1002/sim.1072

Shao J，Chow S-C（2002年）临床试验中的可重复性概率。《医学统计学》21：1727–1742。https://doi.org/10.1002/sim.1177

Siegfried T（2010年）看来，这是错误的。载于：《科学新闻》。
https://www.sciencenews.org/article/odds- are-its-wrong。访问日期：2021年8月22日

Smith CM（2005年）《首先，不伤害》的起源与用途——最重要的是，不造成伤害！《临床药理学杂志》45：371–377。https://doi.org/10.1177/0091270004273680

Smyth GK、Verbyla AP（1999）双重广义线性模型：近似 REML 与诊断。载于：第 14 届国际统计建模研讨会论文集，奥地利格拉茨，第 66–80 页。

Sterne JAC，Smith GD（2001）筛选证据——显著性检验存在什么问题？《物理治疗》81:1464–1469。https://doi.org/10.1093/ptj/81.8.1464

Sterne JAC、Sutton AJ、Ioannidis JPA 等（2011）《随机对照试验荟萃分析中漏斗图不对称性的检验与解释建议》。BMJ 343:d4002。https://doi.org/10.1136/bmj.d4002

Stigler S（2008年）Fisher 与 5%水平。null 21:12–12。
https://doi.org/10.1080/09332480.2008.10722926

Storey JD，Tibshirani R（2003）全基因组研究中的统计显著性。《美国国家科学院院刊》100：9440–9445。https://doi.org/10.1073/pnas.1530509100

Stroebe W，Postmes T，Spears R（2012）科学不端行为与自我神话-科学中的纠正 .
《心理科学视角》7:670–688。https://doi.org/10.1177/1745691612460687 Trafi
mow D、Marks M（2015年）社论。《基础与应用社会心理学》37:1–2。
https://doi.org/10.1080/01973533.2015.1012991 Tsapas A、Avgerinos I、Karagiannis T
等（2020年）2型糖尿病降糖药物的比较疗效。《内科学年鉴》。
https://doi.org/10.7326/M20-0864 Tunç DU、Tunç MN、Lakens D（2021年）二分主张在
统计假设检验中的认识论和语用功能

d

关于统计假设检验

Vallon V、Richter K、Blantz RC 等（1999年）实验性糖尿病中的肾小球高滤过：肾小管重吸收的潜
在作用。JASN 10:2569–2576

Vallon V, Rose M, Gerasimova M 等（2012）敲除钠-葡萄糖转运体SGLT2可减轻糖尿病中
的高血糖和肾小球高滤过，但不能减轻肾脏生长或损伤。《美国生理学杂志：肾脏生理学》
304:F156–F167。https://doi.org/10.1152/ajprenal.00409.2012

Vargas TM, Ferrari SLP, Lemonte AJ（2014）广义线性模型中的改进似然推断。《计算统计
学与数据分析》74:110–124。https://doi.org/10.1016/j.csda.2013.12.002

Wacholder S、Chanock S、Garcia-Closas M 等（2004）评估阳性报告为假报告的概率：一种分子流行病
学研究方法。《JNCI》：《美国国家癌症研究所杂志》96：434–442。
https://doi.org/10.1093/jnci/djh075

Wagenmakers E-J（2007）针对 p 值普遍存在的问题提出一种实用解决方案。《心理学公报与评论》
14:779–804。https://doi.org/10.3758/BF03194105

Wasserstein RL、Schirm AL、Lazar NA（2019）迈向一个超越“p < 0.05”的世界。《美国统
计学家》73:1–19。https://doi.org/10.1080/00031305.2019.1583913

Whitlock MC（2005）合并独立检验所得的概率：加权Z方法优于 Fisher方法。《进化生物学杂
志》18:1368–1373。https://doi.org/10.1111/j.1420- 9101.2005.00917.x

Wilks SS（1932年）方差分析中的若干推广。Biometrika 24:471–494。
https://doi.org/10.2307/2331979

Wood SN（2017年）《广义加性模型：实数集入门》，第二版。Chapman and Hall/CRC，博
卡拉顿

主题索引

索引将是　　在文本定稿后生成，即在外部和内部审查之后　　.