arXiv:2601.05128v1 [stat.ME] 8 Jan 2026

揭示真相：通过高斯求积计算因果推断模拟研究中的真实值

预印本

Alex Ocampo
统计方法学
F. Hoffman La Roche
巴塞尔，瑞士
alex.ocampo@roche.com

Enrico Giudice
高级定量科学
诺华制药股份公司
巴塞尔，瑞士
enrico.giudice@novartis.com

Zachary McCaw
生物统计学系
北卡罗来纳大学教堂山分校
教堂山，北卡罗来纳州，美国
zmccaw@alumni.harvard.edu

Tim P.Morris统计方法学
诺华制药英国有限公司 伦敦，
英国
tim.morris@novartis.com

2026年1月9日

摘要

模拟研究用于了解统计方法的性质。许多模拟研究的一大优势，是能够知道所关注的真实值（即 目标估计量）。掌握了这一预言机知识后，开展模拟研究的研究者便可通过对数据进行反复抽样，评估特定方法恢复真实值的效果。在因果推断模拟研究中，真实值很少是用于生成数据的统计模型中的简单参数。相反，目标估计量通常是平均治疗效应，并对混杂因素和/或中介变量的分布进行边际化。幸运的是，这些变量往往服从正态分布、均匀分布、指数分布或伽马分布等常见分布。对于这些分布，高斯求积能够针对积分核源自已知概率密度函数的被积函数，进行高效且准确的计算。我们通过四个应用示例，展示如何利用高斯求积准确、高效地计算真实因果目标估计量。我们还比较了高斯–埃尔米特求积与蒙特卡洛积分方法的优缺点，并将后者作为基准方法。总体而言，我们证明，高斯求积是一种准确且计算时间可忽略不计的工具，但在模拟研究中仍未被充分用于计算真实因果目标估计量。

1 引言

许多模拟研究涉及从已知分布生成数据，然后考察各种方法恢复参数的能力。研究者对目标参数真实值近乎全知，这是模拟研究相比分析真实数据集能够更深入地了解方法运行特性的原因之一。本文源于我们为因果推断模拟研究生成数据的经验。任何此类分析都必须回答“真实情况是什么”这一问题；也就是说，我们的方法学所针对的因果目标估计量的真实值是多少？在模拟研究中，要公平地评价一种方法，就必须掌握这一信息，以评估偏倚、均方误差和覆盖率等某些基本性能指标。如果真实情况未知或近似得不够准确，那么无偏的方法可能会被误认为存在偏倚，依此类推。

模拟研究通常关注所谓的数据生成模型参数恢复，其推断目标是数据生成模型的某个指定参数，因此其真实值毋庸置疑。在以因果目标估计量为推断目标的模拟研究中，获取真实值往往没那么直接。例如，如果我们根据一个带有 $\beta = 5$的广义线性回归模型进行模拟，并且 β 是我们感兴趣的目标估计量，那么我们知道

我们应使用 β的真实值5来评估偏倚， $\mathbb{E}(\hat{\beta}) - \beta$。在因果推断中，人们关注的是潜在结局分布的函数，而这些函数可能无法由数据生成模型的参数推导得出。例如，适当控制混杂通常十分重要，因此，为了计算我们感兴趣的真实因果效应，必须对所有混杂因素的分布上的函数进行边缘化（或积分）。例如，假设我们感兴趣的是平均因果效应

$$\mathbb{E}[Y^{(a=1)}] - \mathbb{E}[Y^{(a=0)}],$$

但为了针对给定的数据生成机制计算这些期望，我们需要对混杂因素——变量 $C \in \mathbb{R}$ ——的分布进行边缘化，适用于 $Y_{(a)}$。也就是说，对于上述两项，我们都需要执行以下操作[1]:

$$\begin{aligned}
\mathbb{E}[Y^{(a)}] &= \mathbb{E}[\mathbb{E}[Y^{(a)}|C]] \\
&= \mathbb{E}[\mathbb{E}[Y^{(a)}|A = a, C]] \qquad (1) \\
&= \int_{-\infty}^{\infty} \mathbb{E}[Y^{(a)}|A = a, C] f(c)dc, \quad \forall a \in \{0, 1\}. \qquad (2)
\end{aligned}$$

项 $\mathbb{E}[Y_{(a)}|A = a,\ C]$通常被设定为模拟模型的结构方程之一。这些结构方程通常被设定为参数回归模型。这样便可以在给定治疗值为 a 且混杂因素 C 的观测值已知的条件下，模拟（潜在）结局 $Y^{(a)}$ （例如$\mathbb{E}[Y_{(a)}|A = a,\ C] = g^{-1}(\beta_0 + \beta_a a + \beta_C C)$）。然而，即使对于常见的混杂因素分布 C ，当链接函数$g(\cdot)$为非线性时，例如 $g(\cdot)$ = 逻辑斯蒂函数$(\cdot)$，对 C的分布进行积分，通常也无法得到简单的闭式表达式。

解决这一问题的一种方法是从数据生成机制中模拟一个非常大的样本（包括结局），并使用已知无偏的方法进行分析，将所得估计值作为真实值。Morris 等人（2019）审查的 74 项模拟研究中，有 9 项以目标估计量为目标的研究采用了这种基于模拟的方法。其优点有两点：1）对于统计学家而言，只要已知数据生成机制，这种方法就很容易实现；2）大数定律基本上保证了真实值的准确性。不过，这种方法也有两个缺点：首先，由于模拟数据集的规模有限，“真实”值会受到蒙特卡洛误差的影响；其次，生成并分析规模足够大的数据集需要付出计算成本。虽然这并非一个“缺点”，但采用这种方法的前提是：必须存在一种已知对给定数据生成机制无偏的估计量。

最近，Naimi 等人（2025）提出以蒙特卡洛积分作为这一问题更准确且仍具实用性的解决方案。这种方法同样要从数据生成机制中为部分变量生成一个大样本。不过，我们并不模拟观测到的结果，而是生成用于定义目标估计量的潜在结局函数。例如，对于二元结果，我们不抽取一个取值为 0 或 1 的结果，而是根据数据生成模型计算其期望值。我们针对治疗（或更一般地说，感兴趣的原因）的每个取值分别进行计算。最后，计算简单平均值；必要时可对其进行转换，并比较治疗／原因不同水平下的结果。前面提到的结果模拟的两个缺点依然存在，但真实值相关的蒙特卡洛误差更低（在数据集规模相同的情况下）；换言之，使用规模更小的数据集也能达到相同的蒙特卡罗标准误，因此这种方法的计算效率更高。此外，由于这种方法比较的是潜在结局，因此无需具备合适的无偏估计量也可以使用。

避免生成非常大的数据集并接受真实值中的蒙特卡洛误差的一种方法是注意到，在因果推断模拟研究中，我们需要边缘化的变量往往是从共同分布中生成的。例如，考虑式（2）中定义的积分，如果 $C \sim \mathcal{N}(\mu_c,\ \sigma_c^2)$, 这意味着需要计算

$$\int_{-\infty}^{\infty} \mathbb{E}[Y|A = a, C]\phi_C(c)dc, \qquad (3)$$

其中

$$\phi_C(c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{1}{2\sigma_c^2}(c-\mu_c)^2}$$

是 C正态分布的概率密度函数（pdf）。这类积分可以通过高斯–厄米特求积（高斯，1814年；埃尔米特，1864年；Heath，2018年）即时且精确地计算。

本文应用并评估了高斯求积在因果推断模拟研究中计算所关注真实值方面的用途。我们首先概述高斯求积，特别强调高斯–埃尔米特

[1]请注意，式（1）中对 $A = a$ 的条件化在此处是多余的；保留它也无妨，因此为便于读者熟悉，我们仍予以保留。

求积，因为模拟研究中的变量通常由正态分布生成。随后，我们重新考察 Naimi 等人（2025）中的两个示例，并进一步加入两个来自我们自身中介分析模拟研究的示例。最后，我们通过讨论提出建议，说明何时可以考虑采用蒙特卡洛积分或高斯求积方法。我们提供了 R 代码，以促进高斯求积在因果模拟研究中的应用。

2 高斯–埃尔米特求积

2.1 单变量求积

Gauss–Hermite（G–H）求积通过有限和近似以下形式的积分：

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}\,dx \approx \sum_{i=1}^{K} w_i f(x_i). \tag{4}$$

我们将 w_i 称为权重，将 x_i 称为节点。节点是分布在 x 定义域内的取值，经过精心选择，作为 K 次埃尔米特多项式 $H_K(x)$[2] 的根。权重定义为 $w_i = (2^{K-1}K!\sqrt{\pi})/(K^2[H_K(x_i)]^2)$。求和长度 K 决定了求积点的数量，该数量因应用而异。要实现准确的高斯–埃尔米特求积，所需的求积点数量少得出奇；R 软件通常默认使用 $K = 5$。这是因为使用 K 个节点的 G–H 求积能够精确计算 $2K - 1$ 次或更低次数的多项式，其准确度远高于使用黎曼和或梯形近似进行积分。

注重细节的读者可能会注意到，第 1 节中的正态分布变量（例如 $C \sim \mathcal{N}(\mu_c, \sigma_c^2)$）实际上并不符合这种形式。不过，我们可以进行简单的变量替换 $x = (c - \mu_C)/\sqrt{2}\sigma_C$，使其符合正确的形式

$$\begin{aligned}\int_{-\infty}^{\infty} f(c)\phi_C(c)dc &= \int_{-\infty}^{\infty} f(c)\frac{1}{\sqrt{2\pi\sigma_c^2}}\mathrm{e}^{-\frac{1}{2\sigma_c^2}(c-\mu_c)^2}dc \\ &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} f(x\sqrt{2}\sigma_C + \mu_C)\mathrm{e}^{-x^2}dc,\end{aligned}$$

从而得到执行高斯–埃尔米特求积所需的被积函数。因此，无论需要积分的正态变量采用何种参数化，我们都可以应用该求积方法。这可能需要根据 μ_C 和 σ_C 对求积点进行缩放；这是一项简单但重要的工作。在 R 中，可以通过 spatstat.random 软件包进行单变量高斯–埃尔米特求积。

2.2 多元求积

高斯–埃尔米特求积 可扩展至多变量被积函数，以处理多个正态分布需要被 e
考虑的情况。例如，如需全面了解多元高斯–埃尔米特求积，请参阅 Jäckel（2005）。
为了直观理解多元求积，可以考虑需要计算两个积分的双变量情形。例如，

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(c_1, c_2)\phi_{C_1}(c_1)\phi_{C_2}(c_2)dc_1dc_2.$$

例如，如果需要对两个混杂变量进行积分，就可能会遇到这种情况。对于单个变量，我们只需在其定义域内获取求积点；而在双变量情形下，则必须在 C_1 and C_2 的联合定义域内构建求积点网格。

在标准软件中，例如 R 包 mvQuad（Constantin Weiser，2023），需要指定求积的阶数（在单变量情形中对应于 K，例如 5），这将得到 $5^2 = 25$ 个求积点。因此，现在用于近似二重积分的有限和长度为 25。尽管求积点有所增加，多元高斯–埃尔米特求积仍然极为高效，因为计算即使是较大的和通常也很简单。在多元空间中，当变量具有相关性时，通常将求积点以 $\sum_{i=1}^{K}\sum_{j=1}^{K} w_i w_j f$（$c_1$，$c_2$）的方式朴素地放置是

[2]埃尔米特多项式分为两类：概率学家的埃尔米特多项式和物理学家的埃尔米特多项式。高斯求积使用物理学家的埃尔米特多项式，其定义为 $H_K(x) = (-1)^K e^{x^2}\frac{\partial^n}{\partial x^n}e^{-x^2}$。

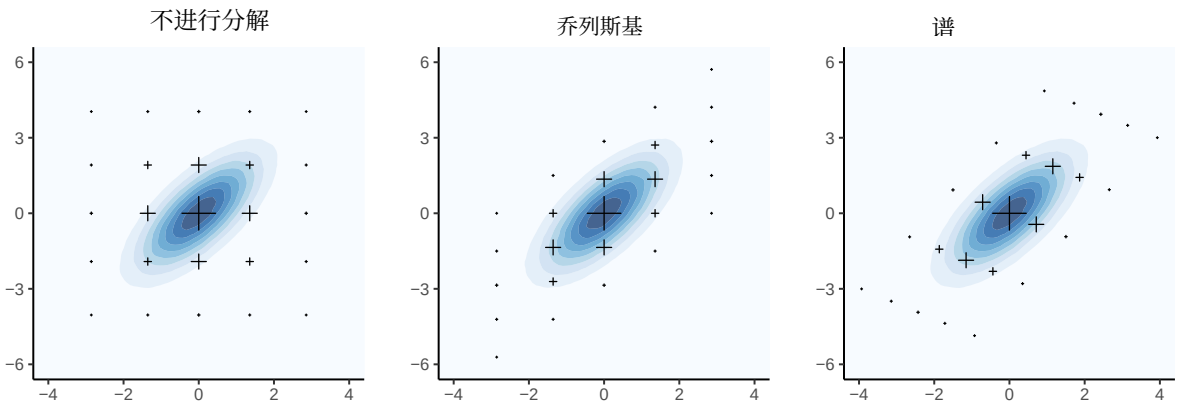

图1：**包含 $K = 5$ 个层级的求积点网格，用于表示二元高斯分布。** 左图：忽略两个变量之间相关性的简单网格。中图：根据 Cholesky分解重新缩放的网格。右图：根据协方差矩阵的谱分解重新缩放的网格。网格点的大小与相应的求积权重成正比。

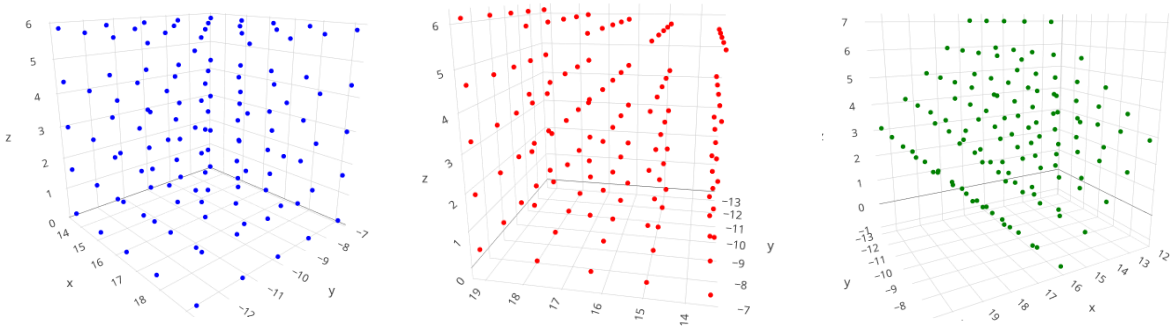

图2：**三维多元高斯分布的求积点三维网格。**左侧的简单网格将（蓝色）求积点均匀放置在网格中，形成一个方形盒。中间的图基于多元正态分布的相关性执行 Cholesky分解，使（红色）求积点略微倾斜。右侧的图根据协方差矩阵的谱分解重新缩放网格，从而形成菱形结构。在这个最终网格中，可以清楚地看出只有两个变量存在相关性，因为在一个维度上，曲面保持不变。

低效的。为克服这一问题，可以采用一种分解方法，根据已知的相关结构旋转求积点网格。将混杂因素之间的相关信息纳入考虑，可以优化求积的精度和计算效率。Jäckel（2005）表明，Cholesky分解比朴素网格更高效，而谱分解的效率更高。这些方法使用以下形式近似积分

$$\sum_{i=1}^{K}\sum_{j=1}^{K} w_i w_j f(c_1, \rho c_1 + \rho' c_2)$$

$$\text{and } \sum_{i=1}^{K}\sum_{j=1}^{K} w_i w_j f(ac_1 + bc_2, bc_1 + ac_2) \text{ respectively,}$$

$$\text{where } \rho = \text{Corr}(c_1, c_2), \quad a = 0.5\left(\sqrt{1+\rho} + \sqrt{1-\rho}\right) \text{ and } b = 0.5\left(\sqrt{1+\rho} - \sqrt{1-\rho}\right).$$

图1直观展示了使用这些分解方法时网格点的布置方式，其中 $K = 5$ 和$\rho = \sqrt{0}$。5。

2.3 其他类型的高斯求积

到目前为止，我们一直关注使用高斯–埃尔米特求积来近似类型（4）的积分，因为模拟情景中的连续变量通常被赋予高斯分布。然而，在某些情况下，非高斯变量可能具有相关性。当模拟研究考察具有不同定义域的分布时，或希望探索放宽高斯分布假设后结果将受到何种影响时，就可能出现此类情景。

高斯–埃尔米特求积是更广泛的一类方法（称为高斯求积方法）的一个特例。高斯求积方法通过在特定点计算函数值，并将其加权求和，从而准确近似特定类型的积分。在模拟研究中，一般的高斯求积方法可用于计算连续随机变量的函数 f 的期望，该随机变量的密度函数为 $p(x)$：

$$\int f(x)\,p(x)\,dx \;\approx\; \sum_{i=1}^{K} w_i\, f(x_i).$$

一般的高斯 求积问题涉及使用次数为 K suc 的正交多项式族 $G_K(x)$ h that

$$\int p(x)\,x^{j}\,G_K(x)\,dx \;=\; 0, \quad \forall\, j = 0, \ldots, n-1.$$

多项式 $G_K(x)$ 的根随后将用作求积点 x_i，而权重 w_i 则由求积对最高 $2K-1$ 次的多项式均为精确这一条件确定。权重和求值点
因此，高斯求积公式的权重和求值点由概率密度 $p(x)$ 唯一确定：对于概率密度 $p(x) \propto e^{-x^2}$ 的情况，对于多项式 $f(x)$，直到 $2K-1$ 阶都能提供精确结果的求值点由高斯–埃尔米特多项式的根给出。表1展示了不同的求积公式，以及相应的密度核；函数正是相对于这些密度核进行隐式积分的。有关与不同函数 $p(x)$ 相关的最优求积所对应的多项式的完整列表，请参见 Zwillinger（2002）。

求积公式	积分类型	隐式核
高斯–厄米特	$\int_{-\infty}^{\infty} \mathrm{e}^{-x^2} f(x)\ dx$	高斯
高斯–勒让德	$\int_{-1}^{1} f\ (x)\ dx$	均匀分布
高斯–拉盖尔	$\int_{0}^{\infty} \mathrm{e}^{-x} f(x)\ dx$	指数分布
广义高斯–拉盖尔	$\int_{0}^{\infty} x^{\alpha} \mathrm{e}^{-x} f(x)\ dx$	伽马分布

表1：不同求积规则、其对应的积分以及相应的概率密度核概述，$f(x)$ 按照这些核进行积分。广义高斯–拉盖尔求积适用于任意实数 $\alpha > -1$。

有关上述求积规则的详情，以及如何利用这些规则对一般非高斯密度进行积分，将在附录 A 中介绍。

3 混杂的应用

3.1 正态分布混杂因素

第一个示例出自 Naimi 等人（2025），研究关注的是边际优势比，其定义如下：

$$\Psi_{OR} \;=\; \frac{\mathbb{P}(Y^{(1)}=1)}{\mathbb{P}(Y^{(1)}=0)} \Big/ \frac{\mathbb{P}(Y^{(0)}=1)}{\mathbb{P}(Y^{(0)}=0)}. \tag{5}$$

为估计这一量，我们对一个单变量正态分布混杂因素 C进行调整。数据生成机制遵循图 3 所示的简单有向无环图结构，其中混杂因素变量服从正态分布，结局服从参数逻辑回归模型：

$$\begin{aligned} C \;&\sim\; \mathcal{N}(\mu_c, \sigma_c^2) \\ \mathbb{P}(Y=1 \mid A, C) \;&=\; \operatorname{expit}(\beta_0 + \beta_1 A + \beta_2 C). \end{aligned} \tag{6}$$

请注意，生成完整的模拟数据集时需要使用暴露的参数模型 A 。然而，由于方程（5）中的潜在结局对暴露进行了干预，因此计算真实值 Ψ_{OR}时无需使用 A 的模型。

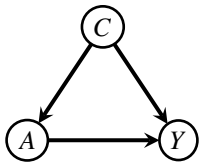

图 3：简单混杂情景的有向无环图。

在此设定下计算因果量，需要求解形式如（4）的积分，以计算 $\mathbb{E}[Y_{(a)}] = \mathbb{P}(Y_{(a)} = 1)$，其中 $a = \{0, 1\}$。在该示例中，随后使用这些量推导边际优势比（5）。
由于难以通过解析方法计算量 $\mathbb{P}(Y^{(a)} = y)$，如 Naimi 等人（2025）所示，可以使用蒙特卡洛积分和数值积分对其进行估计。受计算和复杂度限制，这些方法仅适用于单变量混杂因素 C。在此，我们将这一示例扩展到多变量混杂变量 C，并应用高斯–埃尔米特求积计算真实的边际优势比。

该问题涉及对单个正态分布混杂因素 C进行积分。为简单说明高斯–埃尔米特求积在此情境中的价值，我们将模型 (6) 中的模拟参数 β_0，β_1，σ_c^2 设为 1，$\mu_c = 0$ 和 $\beta_2 = -1$，然后应用具有 $K = 20$ 个求积点的高斯–埃尔米特求积。以下代码片段展示了在 R 中实现该方法的方式：

```
require(mvQuad)
PrY <-function(c,a) 1/(1+exp(c-a-1)) grid <-createNIGrid(dim = 1,
type = "GHN", level = 20) mu1 <-quadrature(PrY, grid, a = 1)
mu0 <-quadrature(PrY, grid, a = 0) (mu1/(1-mu1)) /
(mu0/(1-mu0)) # 优势比
```

可以通过 rescale.NIGrid(m $= \mu_c$, C $= \sigma_c^2$) 函数实现 μ_c 和 σ_c^2 的其他选项。
我们将通过高斯–埃尔米特求积得到的真实值，与通过从数据生成机制中模拟大量 N潜在结局 $Y^{(a)}$ 并取样本均值作为 $\mathbb{P}(Y^{(a)} = 1)$的无偏估计量所得到的估计值进行比较：

$$\begin{gathered}\widehat{\mathbb{P}}(Y^{(a)} = 1) = \frac{1}{N}\sum_{j=1}^{N} Y_j^{(a)}, \\ \text{where } Y_j^{(a)} | C = c_j \sim \text{Bernoulli}(\text{expit}(\beta_0 + \beta_1 a + \beta_2 c_j)) \\ \text{and } c_j \sim \mathcal{N}(\mu_c, \sigma_c^2).\end{gathered} \tag{7}$$

然后，我们将 $\widehat{\mathbb{P}}(Y^{(1)} = 1)$ 和 $\widehat{\mathbb{P}}(Y_{(0)} = 1)$ 代入 (5)，以估计真实值 Ψ_{OR}。值得注意的是，用于确定 Ψ_{OR} 的程序不需要了解分布 fo C 的具体信息（例如正态性），只要能够从中生成样本即可。

此外，我们还使用 Naimi 等人（2025）的蒙特卡洛积分估计真实优势比，即从 C的分布中抽取大量样本，计算潜在结果的期望值：

$$\begin{gathered}\widehat{\mathbb{P}}(Y^{(a)} = 1) \approx \frac{1}{N}\sum_{j=1}^{N} \text{expit}(\beta_0 + \beta_1 a + \beta_2 c_j) \\ \text{where } c_j \sim \mathcal{N}(\mu_c, \sigma_c^2).\end{gathered} \tag{8}$$

同样，可以将 $\widehat{\mathbb{P}}(Y_{(1)} = 1)$ 和 $\widehat{\mathbb{P}}(Y^{(0)} = 1)$ 代入式（5）来计算 Ψ_{OR}。与潜在结果模拟一样，只要能够从 C的分布中生成样本，就无需了解该分布的具体信息。请注意，$\mathbb{P}(Y^{(a)} = 1)$ 的估计值也可用于计算风险差或风险比。

我们省略了简单结果模拟（每行一个结果）作为比较方法。若采用这种方法处理包含 N 个观测值的数据集，其蒙特卡洛误差充其量只能小到与包含 $N/2$ 个观测值的潜在结果模拟相当（其中有两个处理水平）。不等式关系显而易见，因为即使在两种治疗取值下 C 的分布完全相同，并且两种方法在两个取值下模拟的结果数量相同，



图 4：**单个混杂因素下因果优势比的估计值及其 95% 预测区间**。该图比较了基于模拟数据集进行蒙特卡洛积分（标记为“潜在结果模拟”）和对结局的期望值进行蒙特卡洛积分（“蒙特卡洛积分”）所得的平均估计值，以及高斯– 埃尔米特求积估计值。



图 5：**二元混杂因素下因果优势比的估计值及其 95% 预测区间**。该图比较了对结局的期望值进行蒙特卡洛积分（“蒙特卡洛积分”）所得的估计值与高斯–埃尔米特求积估计值。由于潜在结果模拟的误差大得多，会遮挡其他方法之间的比较，因此未将其纳入图中。

潜在结果模拟会在两个处理取值下对相同的 c_i 进行条件化，而简单结果模拟无法做到这一点。

我们绘制了通过高斯–埃尔米特求积和蒙特卡洛积分方法得到的真实边际优势比 Ψ_{OR} 估计值。高斯–埃尔米特估计值基于 $K = 20$ 个求积点。蒙特卡洛积分估计值基于大小为 $N = 10^6$的模拟数据集。由于基于模拟的方法需要模拟 C，因此会受到蒙特卡洛误差的影响；为研究其方差，我们重复了这一过程。因此，我们计算了 $n_{rep} = 10^4$ 个不同的蒙特卡洛估计值，并在图 4 中展示了相应的预测区间。图中还展示了通过 高斯–埃尔米特求积计算的估计值，以及蒙特卡洛估计值的均值。高斯–埃尔米特求积利用了 C已知的高斯分布，因此无需进行模拟，从而避免了任何蒙特卡洛误差。基于模拟结局值的蒙特卡洛估计值，其方差远大于基于期望值的蒙特卡洛估计值。高斯–埃尔米特求积估计值则是确定性的，并且与蒙特卡洛估计均值非常接近。高斯–埃尔米特估计值与蒙特卡洛估计均值之间的绝对差值为 9.51e – 6，相对差值约为 0.0004%。

3.2 多元正态混杂因素

接下来，我们考虑混杂变量的一种更一般的分布，将其定义为由两个相关高斯变量构成的向量：

$$\begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} -5 \\ -10 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \right). \tag{9}$$

系数 β_2 设为 $\boldsymbol{\beta}_2 = [0。1\ 0。1]$。由于两个混杂因素 C_1 和 C_2之间存在相关性，我们按照第 2.2 节所述的步骤，对求积点网格进行谱分解变换。与前一个示例一样，我们将高斯–埃尔米特求积的结果与蒙特卡洛积分估计值进行比较。该估计值通过从 C_1 和 C_2的分布中生成大量样本，并根据公式（8）计算样本均值（“蒙特卡洛积分”）得到。结果如图 5 所示。请注意，由于优势比的不可折叠性（Daniel 等人，2021），Ψ_{OR} 的真实值与第一个示例不同。

高斯–厄米特估计值与蒙特卡洛估计值均值之间的绝对差为 7。59e –7，相对差小于 0。00003%。

3.3 非高斯混杂因素

如第 2.3 节所述，可使用不同的求积规则来计算高斯情形以外的结果。我们考虑三种不同的模拟情景，每种情景都包含一个由两个独立变量组成的二元混杂因素 $\begin{bmatrix} C_1 & C_2 \end{bmatrix}'$：

1. 均匀分布混杂因素： $C_1 \sim \mathrm{U}(-2,\ 2)$， $C_2 \sim \mathrm{U}(-4,\ 0)$。

2. 指数分布混杂因素： $C_1 \sim \mathrm{Exp}(1)$， $C_2 \sim \mathrm{Exp}(2)$。

3. 伽马分布混杂因素： $C_1 \sim \mathrm{Ga}(1,\ 2)$， $C_2 \sim \mathrm{Ga}(4,\ 2)$。

在上述所有模拟设置中，结局概率设定为

$$\mathbb{P}(Y = 1 \mid A, C_1, C_2) \;=\; \mathrm{expit}(-A + 0.5\,C_1 + 0.5\,C_2).$$

我们设计这些情景，是为了得到相应因果优势比的表达式，并能对其进行精确计算。模拟过程的更多细节见附录 A。

我们应用第 3.1 节所述的两种方法：对结果概率 $\mathbb{P}(Y = 1 \mid A,\ C)$进行蒙特卡洛积分抽样（标记为“蒙特卡洛积分”），以及高斯求积。所有蒙特卡洛估计均基于 $N = 10^6$个样本；高斯求积估计值则基于 $K = 20$ 个求积点。对于三种情形中的每一种，均根据表1选择适当的求积公式。结果见图6，其中的直线表示基于 10^4 次运行得到的蒙特卡洛估计均值。蒙特卡洛积分因其方差而在均方误差方面表现较差，而该方差源于此方法的抽样不确定性。高斯求积方法在偏倚方面也优于蒙特卡洛积分：蒙特卡洛估计的平均值虽然基于 10^{10} 个样本，但其偏倚仍高于求积方法（另见图7）。

3.4 收敛与运行时间

最后，我们研究高斯求积如何随着求积点数量的变化而收敛到真实值。如第 2.1 节所述，高斯求积方法对于阶数不超过 $2K - 1$ 的多项式 $f(x)$ 是精确的。对于光滑函数 $f(x)$，高斯求积呈现指数收敛（Barrett，1961），而蒙特卡洛方法则具有恒定且缓慢的收敛速率 $O(N^{-1/2})$。图 7 展示了偏倚绝对值随求积点数量增加而变化的情况。由于在指数分布混杂因素情形下，真实因果优势比具有闭式表达式，因此我们采用高斯–拉盖尔求积。运行时间取 10^5 次运行的平均值。

结果表明，高斯求积的偏倚确实会随着求积点数量的增加而呈指数下降。作为参考，图中以蓝色展示了两个蒙特卡洛积分估计，分别基于 10^6 和 10^8 个样本。基于 7 个求积点的高斯求积，其表现优于基于 10^8 个样本的蒙特卡洛估计。在右侧图中，运行时间似乎会随求积点数量近似线性增加。由于超过 50 个求积点时已接近机器精度，因此不予考虑。未展示蒙特卡洛估计的运行时间，因为其量级分别为 1 秒和 10 秒，比高斯求积分别慢 10^3 和 10^4 倍。

尽管在这些示例中，求积方法明显优于蒙特卡洛方法，但标准求积方法需要在每个待积分维度上分别计算 K 个求积点。因此，这些方法会受到维数灾难的影响，也就是说，计算一个维度为 D 的积分，其复杂度为 $\mathrm{O}(K^D)$。蒙特卡洛方法则能轻松处理高维积分，因为从多元高斯分布中抽样的复杂度为 $\mathrm{O}(D^2)$（Thomas 和 Luk，2008）。研究者提出了能够处理更高维度并生成稀疏网格的求积方法（Heiss 和 Winschel，2008），这些方法不会受到维数灾难的影响。它们基于稀疏构造，能够达到与传统多元网格相当的精度。

图 8 展示了随着混杂变量 C维度增加，高斯–埃尔米特求积和蒙特卡洛方法的运行时间变化。模拟参数取自第 3.2 节的示例。随着维度增加，高斯–埃尔米特求积的运行时间呈指数级增长，而蒙特卡洛方法的运行时间增长较为缓慢。

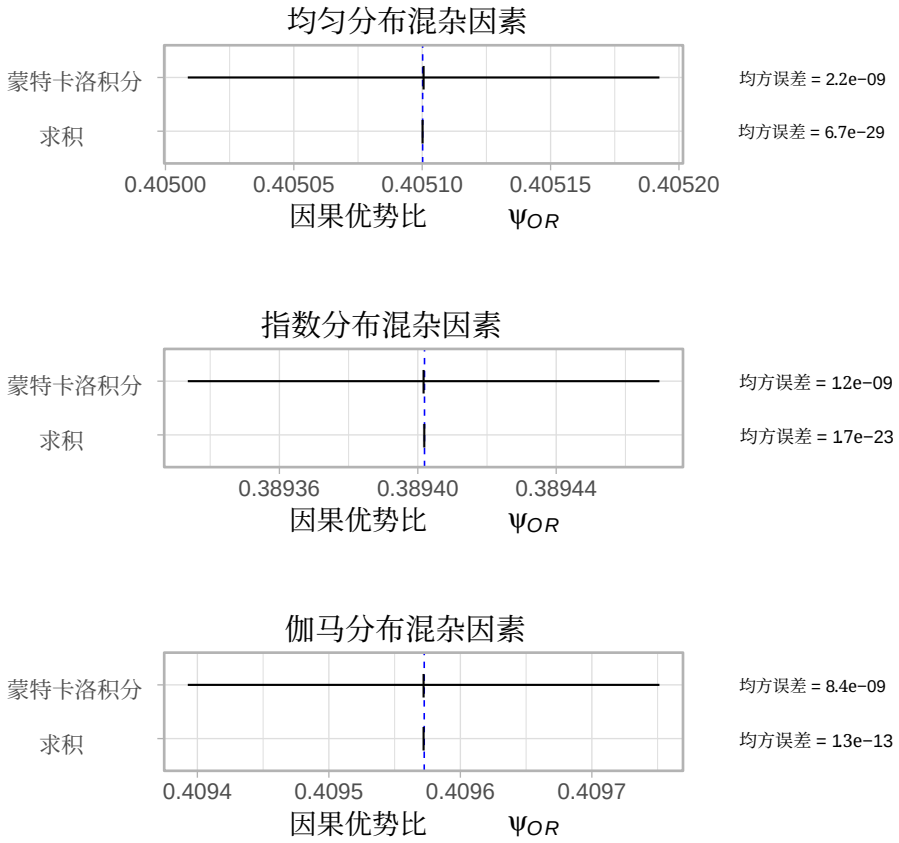

图6：在不同混杂因素分布下，使用两种不同方法估计因果优势比时的因果优势比估计值及其95%预测区间。蓝色虚线表示真实的因果优势比值。

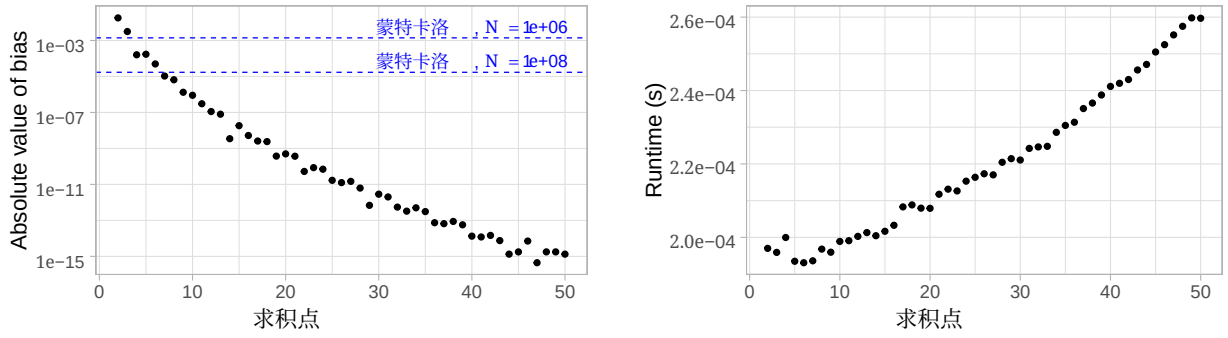

图7：**拉盖尔求积的偏倚和运行时间随求积点数量增加的变化**。左图显示了偏倚（对数尺度）随求积点数量的变化。图中以蓝色显示了两个蒙特卡洛积分估计值作为参考，分别基于 10^6 和 10^8 个样本。右图显示了拉盖尔求积在 10^5 次运行中的平均运行时间。左图中蒙特卡洛估计的运行时间分别约为1秒和10秒。

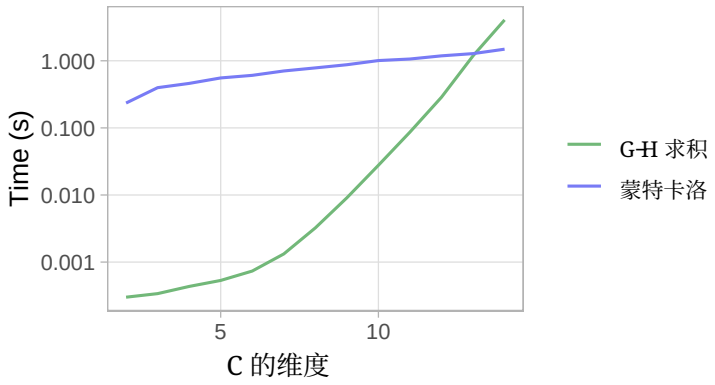

图 8：高斯–埃尔米特求积与蒙特卡洛积分在不同协变量维度下的运行时间。求积网格点数（$K = 3$）和蒙特卡洛样本数（$N = 10^6$）经过选择，以获得相当的偏倚。

4 中介分析的应用

4.1 受控直接效应

接下来，我们重新讨论 Naimi 等人（2025）的第二个示例。这里关注的是受控直接效应，即在将 M 固定为某个值 m 时，治疗 $a = 1$ 与对照 $a = 0$ 下潜在结局的差异。例如，在临床试验中，这可以对应于一个假设性估计目标，即考察如果伴发事件 M 从未发生，平均治疗效应会是多少（欧洲药品管理局，2019年）。受控直接效应的一个例子是：假设患者在试验期间从对照治疗组转为试验治疗组的情况从未发生，治疗效应会如何。这种转换在肿瘤学临床试验中通常是允许的，但绝不会成为未来诊疗的一部分。该目标估计量在数学上可定义为：

$$\Delta_{CDE} = E[Y^{(a,m)}] - E[Y^{(a^*,m)}],$$

其中 $a \neq a^*$，我们将 $Y^{(a,m)}$ 定义为在治疗操作 a 和中介变量值 m 下的潜在结局。沿用上述示例，我们可能关注的是：假设患者未在临床试验中转换治疗组时，治疗的效应（即 $Y_{(a=1,\ m=0)}$ 与 $Y_{(a=0,\ m=0)}$）。模拟的数据生成机制结构由图 9 中的因果有向无环图展示。

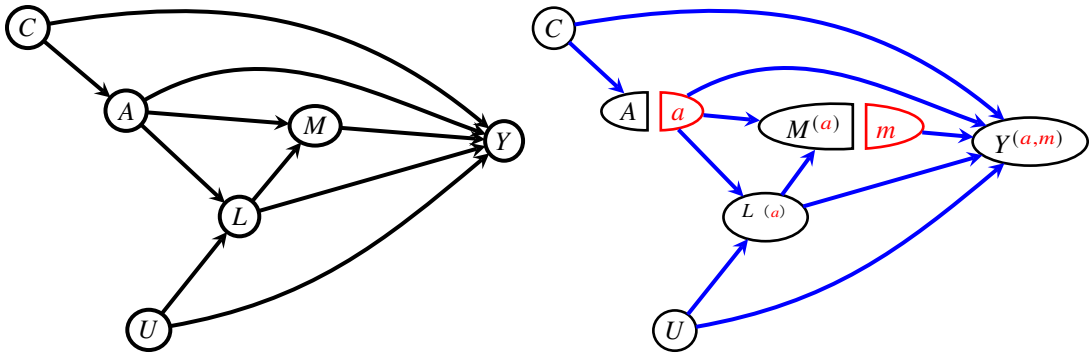

图 9：受控直接效应的因果图。左图：有向无环图。右图：带有干预的 SWIG$A = a$ 和 $M = m$

A t可用于在模拟中生成结局的典型参数回归模型 is

$$\mathbb{E}[Y^{(a,m)}|C, A, L, M, U] = g^{-1}(\beta_0 + \beta_1 a + \beta_2 m + \beta_3 C + \beta_4 L + \beta_5 U),$$

采用某种链接函数 $g(\cdot)$。给定这一数据生成机制（及模型），我们可以使用 g公式（Robins，1986）根据观测数据表示感兴趣的潜在结局：

$$\begin{aligned}\mathbb{E}[Y^{(a,m)}] &= \mathbb{E}_C[\mathbb{E}[Y^{(a,m)}|C]] &(10)\\ &= \mathbb{E}_C[\mathbb{E}[Y^{(a,m)}|C, A=a]] &(11)\\ &= \mathbb{E}_C[\mathbb{E}_L[\mathbb{E}[Y^{(a,m)}|C, A=a, L]]] &(12)\\ &= \mathbb{E}_C[\mathbb{E}_L[\mathbb{E}[Y^{(a,m)}|C, A=a, L, M=m]]] &(13)\\ &= \mathbb{E}_C[\mathbb{E}_L[\mathbb{E}[Y|C, A=a, L, M=m]]] &(14)\end{aligned}$$

g公式中使用的独立性分别是（$Y^{(a,m)} \perp\!\!\!\perp A$）$|C$ 在（11）中调用，以及（$Y^{(a,m)} \perp\!\!\!\perp M$）$|$（$C,\ A=a,\ L$）在（13）中调用；这些独立性可从图 9 中的模板单世界干预图（SWIG）中读出（Richardson 和 Robins，2013；Ocampo 和 Bather，2023）。该图描述了定义受控直接效应时必须考虑的世界。其余等式则分别由对（10）和（12）应用迭代期望，以及在（14）中应用一致性假设（Cole 和 Frangakis，2009）得出。在模拟中，我们还有 U，可以进一步对其积分，得到：

$$\begin{aligned}\mathbb{E}[Y^{(a,m)}] &= \mathbb{E}_C[\mathbb{E}_L[\mathbb{E}_U[\mathbb{E}[Y|C, A, L, M, U]]]]\\ &= \mathbb{E}_C[\mathbb{E}_L[\mathbb{E}_U[g^{-1}(\beta_0 + \beta_1 a + \beta_2 m + \beta_3 C + \beta_4 L + \beta_5 U)]]]\end{aligned}$$

现在假设混杂因素服从正态分布：

$$\begin{aligned}C &\sim \mathcal{N}(\mu_C, \sigma_C^2), \quad \mu_C = -10\\ U &\sim \mathcal{N}(\mu_L, \sigma_L^2), \quad \mu_L = 3\\ L|U &\sim \mathcal{N}(\mu_U, \sigma_U^2), \quad \mu_U = 15 + a + 0.1U\end{aligned}$$

以及 $\sigma_C^2 = \sigma_U^2 = \sigma_L^2 = 1$。请注意，$U$ 是 L 的原因，这会使二者产生相关性。在实际模拟中，我们还需要指定 A 和 M的数据生成机制；但计算真值时无需指定，因为在考虑定义目标估计量的特定潜在结局时，它们被设定为特定值 a 和 m。

根据上述内容，我们可以用以下积分表示前面推导出的 g公式：

$$\mathbb{E}_C[\mathbb{E}_L[\mathbb{E}_U[\mathbb{E}[Y|C, A, L, M, U]]]] = \int_c \int_\ell \int_u \mathbb{E}[Y|C, A, L, M, U]\phi_U(u)du\phi_L(\ell)d\ell\phi_C(c)dc, \tag{15}$$

我们可以使用多元高斯–埃尔米特求积对其进行估计。

式（15）是一个三重积分，因此三元高斯–埃尔米特求积需要在三维空间中进行。每个维度仅使用 5 个节点时，生成的网格包含 5^3 个求积点，因此有限和中共有 125 项。对于这一应用，我们需要使用 Jäckel（2005）的谱分解来纳入 L 与 U 之间的相关性。该分解会根据已知的相关结构旋转网格，从而提高求积的准确性。

从图10可以看出，求积再次准确地估计了真实值。对于连续结局 Y 我们使用了恒等链接，因此可以辨明真实值 12，而求积得出的结果正是该值。蒙特卡罗积分（MCI）的表现也不错，不过存在一定的蒙特卡洛误差，最高达到小数点后三位。我们还模拟了一个真实值未知的二元结果。对于二元结果，求积与蒙特卡罗积分的结果一致，但求积是确定性的。

4.2 应用 3：限制平均生存时间的中介分析

如 Ocampo 等人（2024）所述，一项模拟研究考察了使用伪值（Andersen 和 Perme，2010）对生存时间结局进行中介分析的性质 T。该模拟研究的目标之一，是考察在限制平均生存时间（RMST）尺度上进行中介分析时，该方法的偏倚、均方误差和覆盖率。这些运行特性需要了解真实的中介效应。在本文中，我们重点介绍如何计算所关注的中介估计量的真实值；这一过程需要对中介变量的分布进行积分。

模拟设置如下：令 $A \in \{0,1\}$ 表示治疗组，M 表示连续中介变量的测量。中介变量取自正态分布，其均值取决于 a 的值：

$$M|(A=a) \sim \mathcal{N}(\mu_a, 1), \quad a \in \{0,1\}.$$

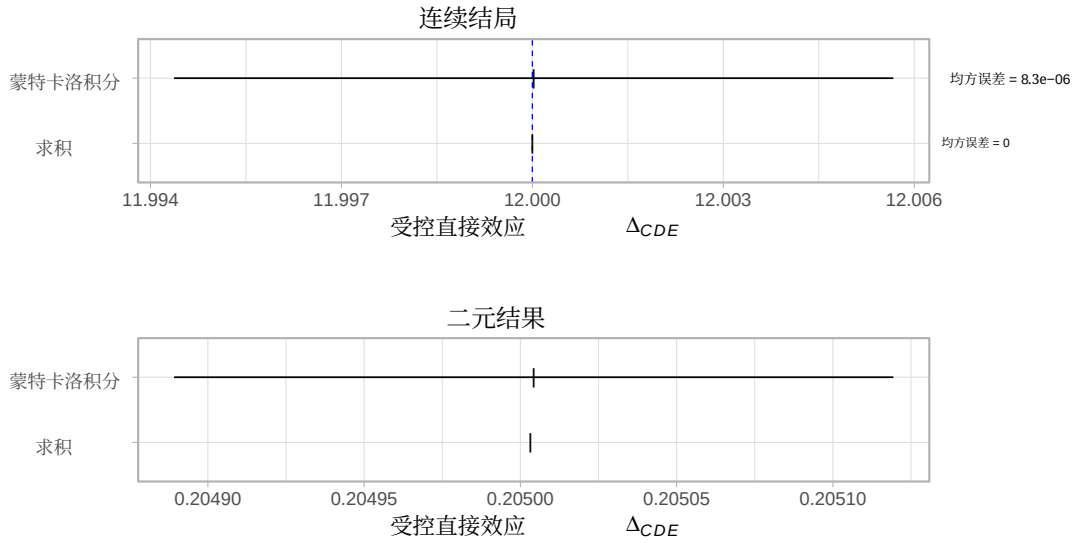

图10：**在受控直接效应模拟中，高斯–埃尔米特求积比蒙特卡洛积分更准确**。蒙特卡洛积分的结果显示了每次样本量为一百万的1,000次重复模拟所得的95%预测区间（10^6）。蒙特卡洛积分的均值是基于从数据生成机制中抽取的1,000个样本（每个样本大小为 10^6 ）估算的均值，并对所有模拟取平均（即等价于使用 10^9 个样本进行蒙特卡罗积分）。蓝色虚线表示在连续结局情形下，根据该模拟设置的数据生成机制得到的真实受控直接效应。

对于数据生成机制，我们定义了 $\mu_0 = 0$ 和 $\mu_1 = -1$。我们假设治疗对中介变量的全部效应发生在任何事件出现之前。然后，从指数分布中模拟事件发生时间（TTE），其参数化形式如下：

$$f(t;\lambda) = \lambda e^{-\lambda t}, \tag{16}$$

其中，率参数 λ 取决于治疗和中介变量：

$$\lambda = \exp\left(\beta_0 + A\beta_A + M\beta_M\right). \tag{17}$$

给定指数分布的随机变量 T，在给定时间 τ 处的真实限制平均生存时间为：

$$\mu(\tau;\lambda) = \int_0^\tau e^{-\lambda t}\,\mathrm{d}t = \frac{1}{\lambda}\left(1 - \mathrm{e}^{-\lambda\tau}\right),$$

据此，我们可以定义总效应、自然直接效应和自然间接效应如下：

$$\mathrm{TE}(\tau) = \mathbb{E}_{M^{(1)}}\left\{\mu(\tau; a=1, M^{(1)})\right\} - \mathbb{E}_{M^{(0)}}\left\{\mu(\tau; a=0, M^{(0)})\right\}, \tag{18}$$

$$\mathrm{NDE}(\tau) = \mathbb{E}_{M^{(0)}}\left\{\mu(\tau; a=1, M^{(0)})\right\} - \mathbb{E}_{M^{(0)}}\left\{\mu(\tau; a=0, M^{(0)})\right\}, \tag{19}$$

$$\mathrm{NIE}(\tau) = \mathbb{E}_{M^{(1)}}\left\{\mu(\tau; a=1, M^{(1)})\right\} - \mathbb{E}_{M^{(0)}}\left\{\mu(\tau; a=1, M^{(0)})\right\}, \tag{20}$$

对于上述所有内容，$M^{(0)} \sim \mathcal{N}(0,\ 1)$ 和 $M^{(1)} \sim \mathcal{N}(-1,\ 1)$。由于所有这些估计目标都以关于正态随机变量（即特定治疗 a 下的中介变量）的期望表示，因此我们可以使用 2.1 节详述的高斯–埃尔米特求积来数值计算真实值。通过求积计算出的 (18)、(19) 和 (20) 的真实值用于说明该方法在覆盖概率适当时不存在偏倚。有关模拟结果的更多详情，请参见 Ocampo 等人（2024）的第 4 节。原始模拟考察了一系列情景和时间点；而本文旨在比较求积法与蒙特卡洛积分，因此我们考察$\tau = 3$。我们按照 Naimi 等人（2025）的做法，在**算法 1**中用伪代码概述了用于限制平均生存时间中介估计目标的蒙特卡罗积分方法。

算法 1 Ps 用于计算真实限制平均生存时间中介估计目标的蒙特卡洛积分伪代码

1：设置随机数生成器的种子值 2：设置较大的样本量 N（例如 $N = 10^9$）

3：模拟来自 $M_{(1)_i} \sim \mathcal{N}(\mu_1, \sigma_1)$ 和 $M_i^{(0)} \sim \mathcal{N}(\mu_0, \sigma_0)$ 的 $i \in 1 \dots N$ 个观测值

4：定义 $\lambda\left(a, M_{(a^*)_i}\right) = \exp\left(\beta_0 + a\beta_A + M_{(a^*)_i}\beta_M\right)$ 并计算

$$\hat{\mu}_i(\tau, a = 1, M_i^{(1)}) = \frac{1}{\lambda(1, M_i^{(1)})}\left(1 - \mathrm{e}^{-\lambda(1, M_i^{(1)})\tau}\right)$$
$$\hat{\mu}_i(\tau, a = 0, M_i^{(0)}) = \frac{1}{\lambda(0, M_i^{(0)})}\left(1 - \mathrm{e}^{-\lambda(0, M_i^{(0)})\tau}\right)$$
$$\hat{\mu}_i(\tau, a = 1, M_i^{(0)}) = \frac{1}{\lambda(1, M^{(0)})}\left(1 - \mathrm{e}^{-\lambda(1, M_i^{(0)})\tau}\right)$$

5：将所有受试者中这些模拟量的均值记为 $i \in 1 \dots N$，即 $\hat{\mu}(\tau, a, M_{(a^*)}) = \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}_i(\tau, a, M_{(a^*)i})$，并用它们计算感兴趣的中介估计量：

$$TE(\tau) = \hat{\mu}(\tau, a = 1, M^{(1)}) - \hat{\mu}(\tau, a = 0, M^{(0)})$$
$$NDE(\tau) = \hat{\mu}(\tau, a = 1, M^{(0)}) - \hat{\mu}(\tau, a = 0, M^{(0)})$$
$$NIE(\tau) = \hat{\mu}(\tau, a = 1, M^{(1)}) - \hat{\mu}(\tau, a = 1, M^{(0)})$$

将高斯-埃尔米特求积与这种使用 10^6 个样本的蒙特卡洛积分方法进行比较时，我们再次发现，高斯求积方法通常更准确（图 11）。1000 次重复蒙特卡洛积分（每次样本量为 10^6）的平均值收敛于求积估计值。此外，蒙特卡洛积分平均耗时 0.44 秒，而求积耗时 0.073 秒。对于一组真实中介估计量而言，差异虽不算特别大，但考虑到原始模拟涉及 324 个真实值，在整体模拟研究中使用求积所节省的计算时间，比单次运行比较中更为显著。

4.3 应用4：风险比的中介分析

考虑另一项以生存时间为终点的中介模拟研究，但这次我们不再关注限制平均生存时间，而是希望在风险比尺度上进行中介分析。我们仍从假设治疗为随机二元变量且在受试者中均匀分配开始，中介变量服从正态分布$M^{(a)} \sim \mathcal{N}(\mu_a,\ 1)$，其中 $\mu_a = \alpha_0 - \alpha_a a$。据此，我们可以从威布尔分布中抽取生存时间结局：

$$T \sim \text{Weibull}(\gamma, \lambda'),$$
$$\text{where } \lambda' = \lambda \exp(\beta_a a + \beta_m m)^{\gamma}.$$

经过一些代数运算，这种威布尔参数化的生存函数可以写为 ：

$$S(t) = \exp\left\{-\left(\frac{t}{\lambda}\right)^{\gamma} \exp(\beta_a a + \beta_m m)\right\}$$

风险函数为：

$$\lambda(t) = \left(\frac{\gamma}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{\gamma-1} \exp(\beta_a a + \beta_m m)$$

利用基本关系 $f(t) = \lambda(t)S(t)$，我们可以将 T 的概率密度函数写为：

$$f(t) = \left(\frac{\gamma}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{\gamma-1} \exp(\beta_a a + \beta_m m) \exp\left\{-\left(\frac{t}{\lambda}\right)^{\gamma} \exp(\beta_a a + \beta_m m)\right\}$$

在这一特定的中介模拟设定中，我们关注定义在风险比尺度上的自然直接效应（NDE）、自然间接效应（NIE）和总效应（TE）：

$$NDE = \frac{\lambda(T(1, M^{(0)})}{\lambda(T(0, M^{(0)})} \tag{21}$$

$$NIE = \frac{\lambda(T(1, M^{(1)})}{\lambda(T(1, M^{(0)})} \tag{22}$$

$$TE = \frac{\lambda(T(1, M^{(1)})}{\lambda(T(0, M^{(0)})} \tag{23}$$

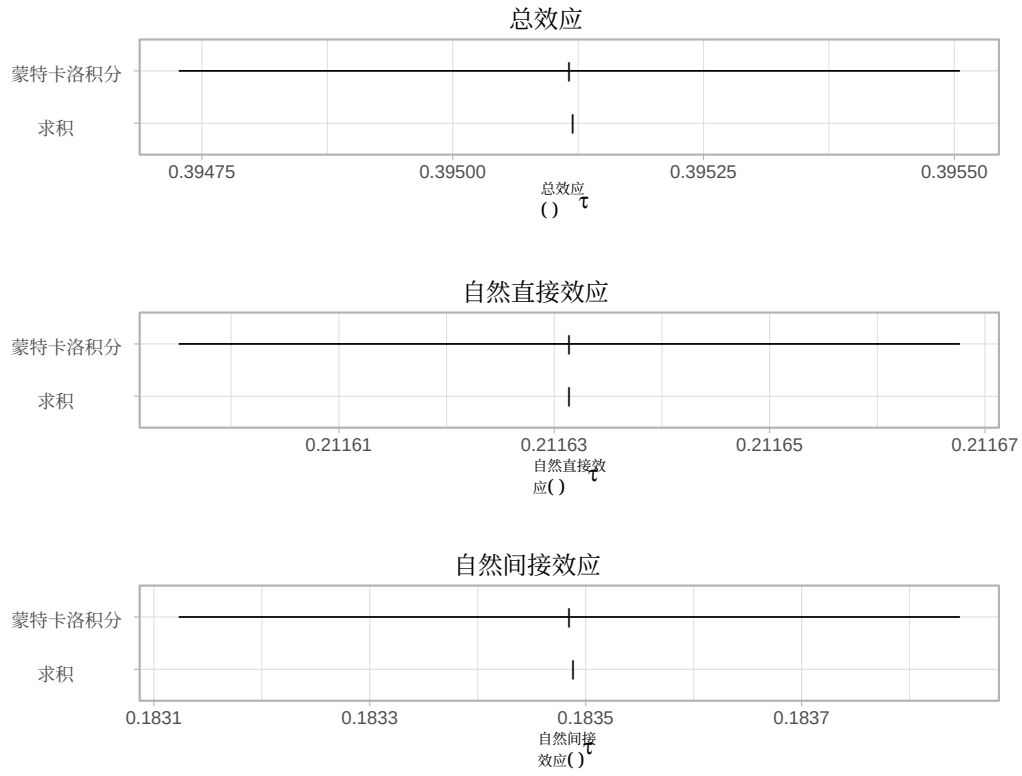

图 11：**在计算真实的限制平均生存时间中介估计目标时，高斯-埃尔米特求积比蒙特卡洛积分更快、更准确**。蒙特卡洛积分的结果展示了数据集大小为 10^6 时，1,000 次重复计算所得的 95% 预测区间。高斯–埃尔米特求积与使用 10^9 个样本的蒙特卡罗积分平均值一致。

这些术语都不对应于用于参数化模拟研究数据生成机制的任何参数（或参数组合）。由于风险率无法表示为期望，我们无法通过对中介变量的分布积分来计算风险率的边际值 $\lambda(t; a,\ m)$。相反，我们利用这样一个事实：$\lambda(t) = {}^{f}(t)\ S\ (t)$，从而计算这些目标估计量的真实值。我们可以使用非参数中介公式（Imai 等人，2010）来确定分子 $f(t)$，

$$f(T_i(a, M^{(a')}) = t) = \int_{\mathcal{M}(a')} f(t_i|A_i = a, M_i = m) dF_{M^{(a')}}(m|A_i = a') \tag{24}$$

$$= \int_{\mathcal{M}(a')} f(t_i|A_i = a, M_i = m)\phi_{M^{(a')}}(m|A_i = a')dm, \tag{25}$$

这可用于推导四种可能的嵌套反事实中任意一种的概率密度函数：$T(1,\ M(1))$、$T(0,\ M(0))$、$T(1,\ M(0))$ 或 $T(0,\ M(1))$；上述中介估计目标正是由这些反事实定义的。由于已知 (21) 中定义的 $f(t)$，且 $M(a)$ 服从正态分布，因此可通过高斯–埃尔米特求积计算。同理，分母 $S(t)$ 可写为：

$$S(T_i(a, M^{(a')}) = t) = \int_{\mathcal{M}(a')} S(t|A_i = a, M_i = m)\phi(m|A_i = a')dm \tag{26}$$

通过迭代期望计算。这也可以用高斯–埃尔米特求积求解。将 (25) 除以 (26)，即可得到计算中介估计目标 (23) 所需的各项。通过计算 t 取一系列合理值时这些泛函的值并求平均，我们便可获得所需的中介效应。

5 讨论

模拟研究是统计学研究中一种极其强大且广泛使用的工具，因为研究者了解数据生成机制，并且能够在一定程度上对其加以控制。从历史上看，模拟研究往往侧重于“参数

恢复”：感兴趣的估计目标通常被方便地定义为数据生成机制的参数。随着人们越来越认识到，估计目标应以非参数方式定义——也就是说，不依赖统计模型——参数恢复不再是无处不在的目标（欧洲药品管理局，2019年）。在开展模拟研究时，制定数据生成机制通常涉及范围确定，而这往往通过试错法完成。借助一种快速计算目标估计量真实值的方法（例如高斯求积），研究者便能更高效地选定所需的模拟参数。准确的真实值至关重要，因为一些模拟性能指标的定义中会用到真实值（见 Morris 等人，2019年，表6）。如果真实值存在微小误差，就会影响估计的性能，使结果看起来存在偏倚、均方误差较高、覆盖不足等问题。

目标估计量并不总能用数据生成机制的参数表示。在这种情况下，我们讨论了以下四种方案：

结果模拟 概念上简单明了，但即使在简单情况下也相对缓慢，且蒙特卡洛误差较大（Morris等，2019）；

潜在结局模拟 **模拟** 与结果模拟类似，但通过同时模拟（ 或全部）潜在逐行数据。

蒙特卡洛积分 与潜在结果模拟类似，但通过计算期望值而非抽取实现值，进一步降低蒙特卡洛误差；这种方法在概念上同样简单，但速度相对较慢（Naimi 等人，2025），且适用于目标估计量为潜在结果均值函数的情况；

求积 在简单情形下不会产生蒙特卡洛误差，且可瞬间完成，但维度较高时可能变慢；只要有足够的求积点 K 并且通过乔列斯基或谱分解考虑相关结构，就能保证精度。

尽管求积具有这些优点，我们也承认它存在一些缺点。首先，在较高维度下，求积仍能保持准确性，但速度会慢于蒙特卡洛积分。其次，求积并非总是适用。例如，由于它利用了关于协变量分布的专门知识，因此无法用于所谓的“plasmode”模拟；这种模拟使用分布未知的观测协变量（Schreck等，2024；Shaw等，2025）。第三，求积需要进行一些代数运算，以求得函数形式并确定被积函数，而结果模拟或蒙特卡洛积分则无需这些步骤。与基于模拟的方法相比，研究人员可能对执行这些计算缺乏信心。在考虑多种数据生成机制时，花时间使用求积可能是值得的，因为求积可以大幅节省时间。此外，由于求积方法具有确定性，其内存效率极高。另一方面，精确的蒙特卡洛估计依赖于存储大量采样数据，因此内存使用量要高得多；这可能很快就会成为问题，尤其是在底层数据生成机制方差较大的情况下。

一个实际问题是，在模拟研究中应在何处报告这一内容。将其归入“数据生成机制”（这是该机制所蕴含的真实值）、“目标估计量”（这是某个目标估计量的真实值），甚至“性能指标”（计算偏倚和覆盖率等指标需要真实值）都说得通。我们认为，“目标估计量”最为合适，但具体细节最终应放在附录或补充材料中；真实值的计算必须正确，但这只是繁琐工作，可能会分散对当前模拟研究其他方面的关注。

致谢

特别感谢诺 华神经科学知识分享会议上的同事提供反馈并参与讨论。

利益冲突

*TPM*持有诺华制药股份公司的股份。他曾从诺华制药股份公司、拜耳医疗保健制药公司、*Alliance Pharmaceuticals*、吉利德科学和*Kite Pharma*获得咨询费，也曾因讲授模拟研究相关内容获得收入。

参考文献

Andersen, P. 和 Perme, M.（2010）。生存分析中的伪观测。《医学研究统计方法》，19(1):71–99。

Barrett, W.（1961）。Co 高斯求积公式的收敛性质。《计算机杂志》，3(4):272–277 .

Cole, S. R. 和 Frangakis, C. E.（2009）。因果推断中的一致性陈述：定义还是假设？《流行病学》，20(1):3–5。Constantin Weiser（2023）。*mvQuad*：多元求积方法。（R 包版本 1.0-8）。Daniel, R.、Zhang, J. 和 Farewell, D.（2021）。化橘为橙：比较针对不同协变量集进行调整后不可折叠效应估计量及其标准误。《生物计量学杂志》，63(3):528–557。欧洲药品管理局（2019）。《临床试验统计学原则指南》的临床试验估计目标和敏感性分析补充文件，ICH E9(R1)。Frederick Novomestky（2022）。*gaussquad*：高斯求积函数集。（R 包版本 1.0-3）。Gauss, C. F.（1814年）。通过近似求出积分值的新方法。Heath, M.（2018年）。科学计算。工业与应用数学学会，第2版。Heiss, F. 和 Winschel, V.（2008）。利用稀疏网格进行数值积分的似然近似。《计量经济学杂志》，144(1):62–80。Hermite, M.（1864年）。关于函数的一种新级数展开。Gauthier-Villars 印刷所。Imai, K.、Keele, L. 和 Tingley, D.（2010）。因果中介分析的一般方法。《心理学方法》，15(4):309。Jäckel, P.（2005）。关于多元高斯–埃尔米特求积的说明。伦敦。荷兰银行。Morris, T. P.、White, I. R. 和 Crowther, M. J.（2019）。利用模拟研究评估统计方法。《医学统计学》，38(11):2074–2102。Naimi, A.、Benkeser, D. 和 Rudolph, J.（2025年）。使用蒙特卡洛积分计算模拟研究中的真实参数值。《流行病学》，36:690–693。Ocampo, A. 和 Bather, J. R.（2023）。用于定义、识别和传达临床试验估计目标的单世界干预图。《医学统计学》，42(21):3892–3902。Ocampo, A.、Giudice, E.、Häring, D. A.、Magnusson, B.、Lange, T. 和 McCaw, Z. R.（2024）。使用伪值简化生存结局的因果中介分析。*arXiv* 预印本 *arXiv:2411.17533*。Rabinowitz, P. 和 Weiss, G.（1959）。用于数值计算形如$\int_0^\infty e^{-x}x^n f(x)dx$的积分的横坐标和权重表。《数学表与其他计算辅助资料》，13(68):285–294。Richardson, T. S. 和 Robins, J. M.（2013）。单世界干预图（SWIG）：反事实方法与因果关系图形方法的统一。统计与社会科学中心、华盛顿大学系列。工作论文，128(30):2013。Robins, J.（1986）。死亡率研究中持续暴露期因果推断的新方法——应用于控制健康工人幸存者效应。《数学建模》，7(9-12):1393–1512。Schreck, N.、Slynko, A.、Saadati, M. 和 Benner, A.（2024）。统计等离子体模拟：潜力、挑战与建议。《医学统计学》，43(9):1804–1825。Shaw, P. A.、Gruber, S.、Williamson, B. D.、Desai, R.、Shortreed, S. M.、Krakauer, C.、Nelson, J. C. 和 van der Laan, M. J.（2025年）。因果推断情境下等离子体模拟研究的注意事项。Thomas, D. B. 和 Luk, W.（2008）。面向可重构硬件的多元高斯随机数生成。《*ACM* 可重构技术与系统汇刊》，1(2):1–29。Zwillinger, D.，编者（2002）。*CRC* 标准数学表与公式。Chapman and Hall/CRC，第31版。

非高斯分布的求积公式

本节探讨高斯–厄米特情形以外的不同求积类型，用于估计形如 $\int f(x)\ p(x)\ dx$ 的期望，其中 $p(x)$ 是相对于 $f(x)$ 进行积分的某个（可能为多元）变量的概率密度。正文中的表1介绍了三种求积类型，每种类型分别对应一种不同的概率密度 $p(x)$。我们详细说明了每种求积公式的实现方法，以便在模拟研究中估计因果量。

均匀分布变量

在模拟研究中，我们需要对其进行积分的已测量/未测量混杂因素可能服从均匀分布。对于区间 $[-1,1]$ 上的均匀概率密度 $p(x) \propto 1$，最优的高斯求积节点x_i 由勒让德多项式 $P_n(x)$ 的根给出。根据高斯求积的构造方式，计算权重是为了使多项式 $f(x)$ 在最高达到 $2n-1$ 阶时能够得到精确结果。相应的权重 w_i 为

$$w_i = \frac{2}{(1-x_i^2)\,[P_n'(x_i)]^2}.$$

求积点及其对应权重专为形如 $\int_{-1}^{1} f(x)dx$ 的积分设计。对于区间 $[a,b]$ 上的一般均匀概率密度，该积分可通过简单的变量替换 经由 $t = \frac{2x-b-a}{b-a}$ 重写：

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}t + \frac{b+a}{2}\right)dt.$$

在第 3.3 节所述的模拟设置中计算因果优势比，需要计算以下积分，其中 $a = \{0,1\}$：

$$\mathbb{P}(Y^{(a)} = 1) = \int_0^\infty\int_0^\infty \frac{1}{16}\,\text{expit}\left(\frac{c_1}{2} + \frac{c_2}{2} - a\right)dc_1dc_2.$$

精确结果可以用二重对数函数 $\text{Li}_2(z)$ 表示，并且可以通过任意精度计算：

$$\mathbb{P}(Y^{(0)} = 1) = \frac{1}{4}\left[2\,\text{Li}_2(-e^{-1}) - \text{Li}_2(-e) - \text{Li}_2(-e^{-3})\right] \quad \text{and}$$
$$\mathbb{P}(Y^{(1)} = 1) = \frac{1}{8}\left[4\,\text{Li}_2(-e^{-2}) - 2\,\text{Li}_2(-e^{-4}) + \frac{\pi^2}{6}\right].$$

指数分布变量

对于指数概率密度 $p(x) = e^{-x}$，最优高斯求积由拉盖尔多项式 $L_n(x)$ 给出。具体而言，每个求积点 x_i 是拉盖尔多项式 $L_n(x)$ 的第 i 个根，相应的权重 w_i 为

$$w_i = \frac{1}{x_i\,[L_n'(x_i)]^2}.$$

求积点及其对应权重专为形如 $\int_0^\infty e^{-x} f(x)dx$ 的积分设计。对于一般指数概率密度 $p(x) = \lambda e^{-\lambda x}$，该积分可通过简单的变量替换经由 $t = \lambda x$重写：

$$\int_0^\infty f(x)\,\lambda e^{-\lambda x}\,dx = \int_0^\infty f\left(\frac{t}{\lambda}\right)e^{-t}\,dt.$$

计算第 3.3 节所述模拟设置中的因果优势比，需要计算以下积分，其中 $a = \{0,\ 1\}$：

$$\mathbb{P}(Y^{(a)} = 1) = \int_0^\infty\int_0^\infty 2e^{-2c_2}\,e^{-c_1}\,\text{expit}\left(\frac{c_1}{2} + \frac{c_2}{2} - a\right)dc_1dc_2.$$

上述积分具有闭式表达式，其精确结果为

$$\mathbb{P}(Y^{(0)} = 1) = \frac{2}{3} \quad \text{and} \quad \mathbb{P}(Y^{(1)} = 1) = \frac{4}{e}\left[\frac{2}{3} + \frac{1}{e}\left(\frac{1}{2} - \frac{1}{e} + \frac{1-e^2}{e^2}\log(1+e)\right)\right].$$

伽马分布变量

对于具有伽马核 $p(x) \propto x^{\alpha} e^{-x}$ 的密度，可以使用更一般的求积公式，通过广义拉盖尔多项式 $L_n^{(\alpha)}(x)$（Rabinowitz 和 Weiss，1959）得到。相应的权重 w_i 为

$$w_i = \frac{x_i \, \Gamma(n+\alpha+1)}{n! \, (n+1)^2 \, [L_{n+1}^{(\alpha)}(x_i)]^2}.$$

求积点及其对应的权重专为形如 $\int_0^{\infty} x^{\alpha} e^{-x} f(x) dx$ 的积分而设计。对于一般的伽马密度 $p(x) \propto x^{\alpha} e^{-\lambda x}$，可通过相同的变量替换 $t = \lambda x$ 将积分改写为：

$$\int_0^{\infty} f(x) \, \frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)} x^{\alpha} e^{-\lambda x} \, dx = \frac{1}{\Gamma(\alpha+1)} \int_0^{\infty} f\left(\frac{t}{\lambda}\right) t^{\alpha} \, e^{-t} \, dt.$$

在第 3.3 节所述的模拟设置中，使用 $C_1 \sim \mathrm{Ga}(1,\ 2)$ 和 $C_2 \sim \mathrm{Ga}(4,\ 2)$ 计算因果优势比，需要计算以下 $a = \{0,\ 1\}$ 的积分：

$$\mathbb{P}(Y^{(a)} = 1) = \int_0^{\infty} \int_0^{\infty} \frac{2^5}{\Gamma(4)} \, e^{-2c_1} \, c_2^3 \, e^{-2c_2} \, \mathrm{expit}\left(\frac{c_1}{2} + \frac{c_2}{2} - a\right) dc_1 dc_2.$$

为了得到一个闭式形式 m 表达式中，利用以下事实可简化积分：$Z = 0。5\, C_1 + 0。5\, C_2 \sim \mathrm{Ga}(5,\ 1)$：

$$\mathbb{P}(Y^{(a)} = 1) = \int_0^{\infty} \frac{1}{\Gamma(5)} \, z^4 \, e^{-z} \, \mathrm{expit}\,(z - a) \; dz.$$

因此，精确结果为

$$\mathbb{P}(Y^{(0)} = 1) = \frac{15}{16} \zeta(5) \quad \text{and} \quad \mathbb{P}(Y^{(1)} = 1) = \frac{3 + 10\,\pi^2 + 7\,\pi^4 - 360\,\mathrm{Li}_5\left(-e^{-1}\right)}{360\, e}$$

其中，$\mathrm{Li}_5(z)$ 是以 5 为底的多对数函数，ζ 是黎曼ζ函数，两者均可通过任意精度计算。

上述结果的多变量扩展很直接，mvQuad R 软件包（Constantin Weiser，2023）提供了多变量高斯–勒让德和高斯–拉盖尔求积规则的实现。对于广义高斯–拉盖尔求积，gaussquad R 软件包（Frederick Novomestky，2022）提供了计算权重和求积点的函数。