一种简单的贝叶斯模型：利用具有层级误差结构的计数数据估计比例和比值，并应用于微滴数字 PCR 实验

Elyas Mouhou* Vincent Audigier† Josselin Noirel*‡

2023年5月5日

arXiv:2305.02700v1 [stat.AP] 4 May 2023

摘要

在许多生物医学领域中，具有层级结构误差的实验设计十分普遍；为了考虑离散度并从数据中获得可靠的推断，纳入这种层级架构至关重要。本文研究了这样一个问题：当生物学重复或技术重复的数量有限时，如何根据类似于液滴数字聚合酶链式反应实验所生成的阳性或阴性计数数据，估计比例或比值。我们提出并讨论了一个贝叶斯框架，为此在 R 中提供并实现了一个吉布斯采样器，并将其与随机效应模型进行比较。

目录

1 引言 **3**

2 方法 2.1 二项近似.4 2.2 近似吉布斯采样器的实现.5

3 ddPCR 数据的应用 **5**

3.1 等位基因比率.5 3.2 测量低肿瘤比例.5 3.3 处理细胞系中的差异表达.6

4 贝叶斯方法比随机效应模型更可靠 7 5 讨论 8 6 代码可用性 9 7 致谢与作者贡献 10

*Laboratoire GBCM (EA7528), 法国国立工艺学院，HESAM Université，2, rue Conté, 75003巴黎 法国。

†CEDRIC，法国国立工艺学院，HESAM Université，法国巴黎孔泰街 2 号，邮编 75003。
‡通讯作者。

A 基于众数近似值的上界 10 B 基于泊松分布的模型设定 10

1 引言

常规估计和假设检验在具有层级误差结构的数据背景下几乎没有相关性。然而，这种情况在生物科学、临床科学和生物技术研究中十分普遍；在这些领域，数据包含两个（有时更多[1]个）层级不同的不确定性来源：生物变异和技术变异。当存在层级误差结构时，Student 的 t 检验及类似方法都必须弃用，而应采用特殊技术。其中一些情况可以使用混合效应模型处理，其目标实际上是估计较高层级的参数。审慎构建的层级模型有助于提高估计结果的稳健性，并增强结果的可信度。

液滴数字聚合酶链式反应（ddPCR）是一种将DNA片段分散到独立液滴中的技术，以便对分离的DNA分子，或至少数量非常有限的DNA分子进行扩增。这使研究人员能够基于近乎单分子分辨率的瞬时采样结果进行定量，无论是确定等位基因频率/等位基因比率（常用于癌症研究），还是确定某个目标基因的表达倍数变化（用于靶向功能基因组学）。图1说明了ddPCR的工作原理。这一实验程序具有以下特点：（1）我们关注的是一个比值超参数；（2）观测数据是具有潜在泊松基础的计数数据。

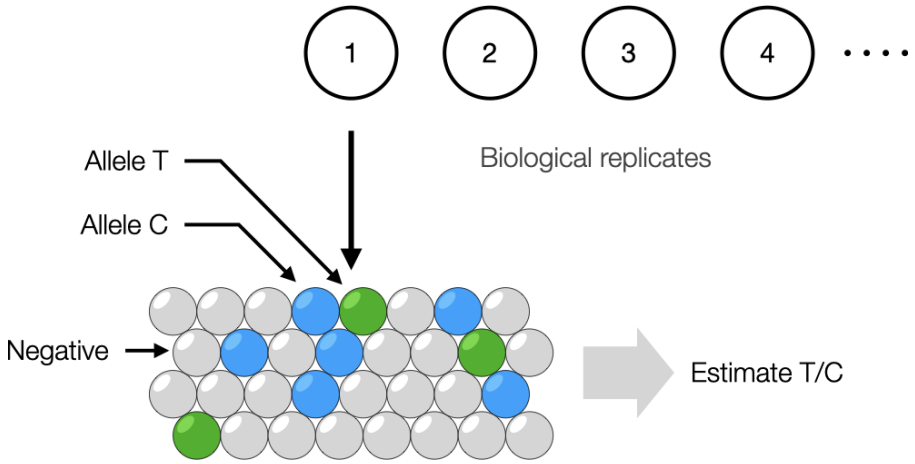


图 1：ddPCR 是一种核酸定量方法，先对样本进行稀释，再将其分配到数千个微小液滴中。每个液滴都充当一个微型反应室，并在其中独立进行 PCR 扩增。

我们通过一个（不完全是）假设情境来说明这一点：开展 k 生物学重复，以估计等位基因比率 $r = A/B$ 或等位基因频率 $f = A/(A + B)$，例如借此了解肿瘤对新治疗 [2]的响应。我们预期，每个生物学重复的（未知）等位基因频率 p_i 会围绕群体总体值f分散。对于每个生物学重复，ddPCR 数据会记录以下信息 i：(1) N_i，即未发生扩增的液滴数（阴性液滴）；(2) A_i，即仅扩增等位基因 A 的液滴数（A 阳性）；(3) B_i，即仅扩增等位基因 B 的液滴数（B 阳性）；以及 (4) D_i，即两个等位基因 A 和 B 均发生扩增的液滴数（双阳性）。表 1 展示了 ddPCR 数据示例。

本文介绍了一种处理 ddPCR 数据点的简单贝叶斯方法，并提供了一个吉布斯采样器（打包在一个 R程序包中，参见第 9 页的代码可用性部分），该采样器

[1]例如，定量蛋白质组学会产生定量数据，其离散度来源于生物学层面、实验层面和技术层面——肽和蛋白质层面——参见 [3, 4]。

将帮助用户处理数字液滴 PCR 数据。

液滴:	C 阳性	T 阳性	双阳性	阴性	总计
重复 1	186	166	1	14 036	14 389
重复 2	159	124	1	14 773	15 057
重复 3	6337	6054	2474	738	15 603

表 1：ddPCR 结果示例。

2 方法

2.1 二项近似

下面介绍用于 ddPCR 数据的层次模型，如图2所示。

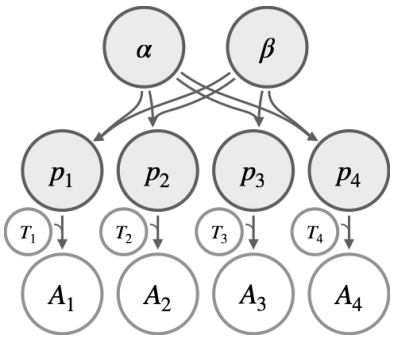


图 2: 用于对 ddPCR 数据进行贝叶斯推断的简单层次模型 .

假设 A 等位基因的比例服从贝塔分布 $B(\alpha, \beta)$，两个超参数α 和 β 控制着生物学重复间的中心趋势和离散度。最终，需要估计参数 α 和 β，以得到等位基因比率 α/β的估计值。因此，生物学重复样本 i 中的 A 阳性液滴比例 p_i 取自贝塔分布 Beta(α, β)； B 阳性液滴的比例记为 $q_i = 1 - p_i$。

为了便于实现该模型，(1) 我们忽略通常可以忽略不计的双阳性液滴 N_i，(2) 假设每个阳性液滴都含有一个独特的 DNA 分子，且 (3) 假设阳性液滴总数 $T_i = A_i + B_i$ 是一个观测常数。由此可知， A_i 服从参数为 Binom(p_i, T_i) 的二项分布。请注意，本文将在末尾讨论这些工作假设（参见讨论，第 8 页）。

为了完整地指定模型，我们对 α 和 β采用不正规无信息先验：假设这两个参数都服从 (1, $+\infty$) 上的均匀分布。这确保了在最坏情况下，比例p_i服从 (0,1) 上的均匀分布。实践中，这种无信息先验既有效，也能为解释各种情况提供稳健的基础：

- 当肿瘤细胞含量较低，且新生突变的比值预计较低时，
- 当在表现出单核苷酸多态性的特定基因座上，体细胞为杂合状态，因此预计比值接近 1 时，
- 在差异表达测量中，比值几乎可以是任意值。

2.2 近似吉布斯采样器的实现

我们的最终模型如下:

$$\alpha \sim \mathrm{Unif}(1, +\infty) \tag{1a}$$
$$\beta \sim \mathrm{Unif}(1, +\infty) \tag{1b}$$
$$p_i \sim \mathrm{Beta}(\alpha, \beta) \tag{1c}$$
$$A_i \sim \mathrm{Binom}(p_i, T_i) \tag{1d}$$

Imp实现吉布斯采样器意味着我们需要写出条件概率 :

$$P(p_i \mid \alpha, \beta) \sim \mathrm{Beta}(\alpha + A_i, \beta + B_i) \tag{2a}$$
$$P(\alpha \mid \beta, p_i) \propto \mathrm{B}(\alpha, \beta)^{-k} \times P^{\alpha} \tag{2b}$$
$$P(\beta \mid \alpha, p_i) \propto \mathrm{B}(\alpha, \beta)^{-k} \times Q^{\beta} \tag{2c}$$

其中，B(a，b) 是贝塔函数，K 是重复样本的数量，P 是乘积 $P = \prod p_i (1 \le i \le k)$，而 Q 是乘积 $Q = \prod q_i$ $(1 \le i \le k)$。这里的主要困难在于，据我们所知，出现在式 (2b) 中并由 β、k 和 P 参数化的概率密度函数如下（C 是归一化因子）

$$P(\alpha; \beta, k, P) \mapsto C \times \mathrm{B}(\alpha, \beta)^{-k} \times P^{\alpha} \quad \alpha > 1 \tag{3}$$

并不是一种广为人知的概率分布。（式 (2c) 也同样如此。）为了使采样器尽可能简单，我们提出以下解决方案：将 α 的取值离散化，并限制为大于或等于 1 的整数；此外，我们设置一个上界，即某个整数 M，可以手动设定，也可以使用数值启发式方法来确定（见附录 A）。在实践中，可以假设上界为 $M = 1000$，并在事后检查从该分布中采样得到的值是否远小于 M。

3 ddPCR 数据的应用

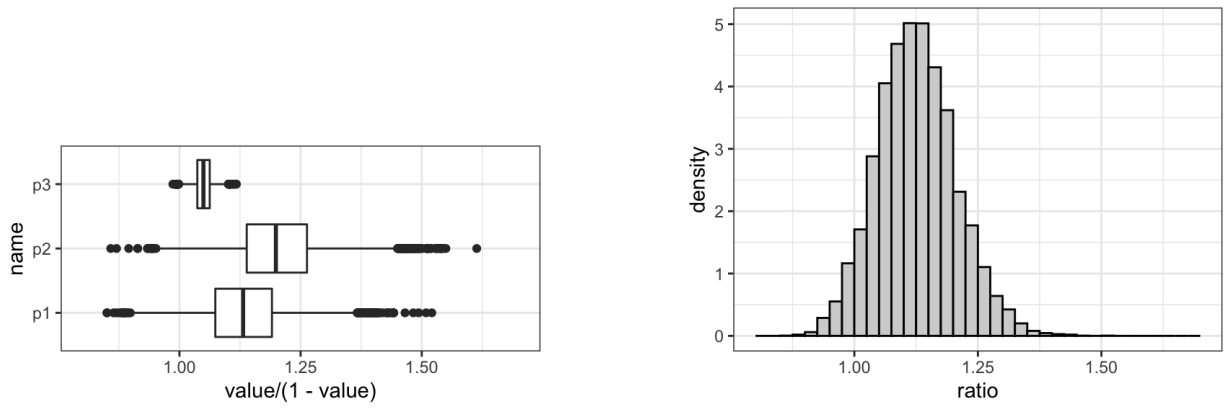
3.1 等位基因比率

为根据表 1 中的数据估计 C/T，我们使用以下参数运行吉布斯采样器：上界为 $M = 5000$，预热迭代次数为 100，抽稀因子为 25，从而获得 10 000 个样本；据此可推导出 α、β 以及各个 p_i 的后验分布。我们可以聚焦于单个重复样本，求得各个等位基因比率的可信区间 $p_i/(1 - p_i)$（参见图 3A）；也可以着眼于整体，即理想化等位基因比率 α/β 的抽样值（参见图 3B）。

在此使用贝叶斯框架的优势在于，可以针对等位基因比率作出概率判断；例如，假设模型的一般贝叶斯假设大体成立，我们可以计算比率明显低于某个有意义阈值的概率。我们来看看表 1 中给出并在此分析的例子：比率 C/T 是否可能低于 1？吉布斯采样器给出的答案是，概率 Pr($r < 1 - \epsilon$) 约为 3%（取值 $\epsilon = .01$，这一选择有些随意）。这可能与基于各个 p_i 的点估计进行的朴素 Student t 检验所呈现的情况不同（在这里，该检验会得出一个平平无奇的 p 值，约为 $P = .09$）。

3.2 测量低肿瘤比例

基于液滴的数字 PCR 通常用于根据肿瘤特异性体细胞突变（$GNAQ$ Q209R）估算肿瘤比例。下方的数据集旨在评估



A： p_i的箱线图　　B： β/α 的直方图

图 3：各个独立重复样本的 C/T 比值后验分布（以箱线图表示）以及总体理想化 C/T 比值的后验分布（以直方图表示）。

在居里研究所开展的一项实验中，将携带该突变的肿瘤 DNA 与健康供体的基因组 DNA 混合稀释，以评估 ddPCR 的技术灵敏度。

液滴:	C 阳性	T 阳性	双阳性	阴性	总计
重复 1	9	2245	1	10 279	12 534
重复 2	4	2414	0	10 258	12 676
重复 3	3	2102	1	9386	11 492
重复 4	1	2449	1	11 156	13 607
重复 5	2	2739	2	12 578	15 321
重复 6	4	2442	1	12 158	14 605

表 2：等位基因比率为 C $\ll$ T 的 ddPCR 数据示例。

这个例子很有意思，因为它展示了两种等位基因浓度极不平衡的情况；因此，点估计噪声较大，采用规范统计方法进行稳健推断显得至关重要。

模型对总体比率的点估计为 0.002，95%可信区间为 (0.001,0.003)。请注意，在这个例子中，值得检查上界 M 是否远高于 α 和 β 的采样值（$M = 4000$ 是令人满意的）。

3.3 处理细胞系中的差异表达

如果我们想了解某个等位基因（此处为在细胞系（siNKX2.4）中引入的等位基因）导致的基因表达水平变化（此处的基因为 $CLPTM1L$] ），就必须将该基因的表达量与某个参考基因（此处为 $GUSB1$）的表达量进行比较。ddPCR 技术用于扩增 $CLPTM1L$ 或 $GUSB1$扩增子，而不是等位基因。可以用 α/β 比值来估算$CLPTM1L$ 相对于$GUSB1$的转录水平。因此，转录水平的变化将以比值之比来量化

$$\frac{\{\alpha/\beta\}_{\text{in siNKX2.4}}}{\{\alpha/\beta\}_{\text{in CTRL}}} \tag{4}$$

我们需要处理的数据可能类似于表 3 所示的数据。

如果假设针对各个细胞系收集的数据彼此独立，我们便可以直接根据分别从两种细胞系独立获取的样本，评估表达量相对变化 (4)

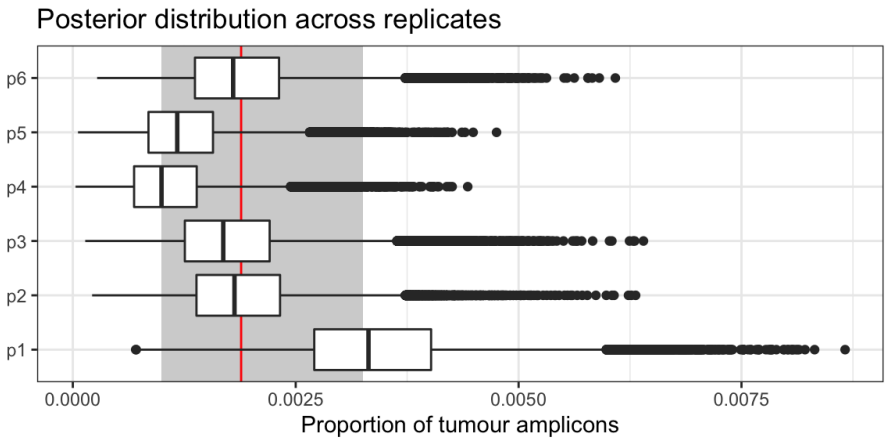


图4： p_i的后验估计（以箱线图表示）以及总体比率（阴影区域表示95%可信区间，红线对应点估计）。

	siCTRL细胞系		siNKX2.4细胞系	
液滴：	CLPTM1L	GUSB1	CLPTM1L	GUSB1
重复 1	1905	841	2604	1063
重复 2	1640	742	1625	540
重复 3	1829	835	1550	600
重复 4	1581	811	1793	650
重复 5	3819	2139	4538	2243

表 3：差异基因表达数据示例。

如图 5 所示。随后利用样本根据公式 (4) 计算表达量相对变化，并据此获得统计量。应用于表 3 中的具体数据，可得表达量相对增加的点估计为 1。23，95% 区间为：(1。04，1。45)。

4 贝叶斯似乎比随机效应模型更可靠

具有层级结构的比例可以通过逻辑回归进行建模，其中每个重复级比例对应一个随机效应。

我们使用贝叶斯模型分析了 Alan Agresti 的一个数据集，并将其估计结果与随机效应模型 [1] 的估计结果进行了比较。

表 13.2（第499页）第13章“聚类分类数据：随机效应模型”第13.3节“二元数据的随机效应模型示例”

该分析旨在根据州级计数估计奥巴马的得票比例（该数据集实际上是虚构的）。结果如图6所示。这里的要点是，随机效应估计结果与贝叶斯估计结果一致，并且在除五个州以外的所有州中，95%可信区间都与真实值相容。总体置信区间和可信区间具有可比性，分别为（0.466，0.549）和（0.456，0.548）。

表 1 提供了另一个有趣的测试案例，因为生物学重复数量较少。本文所述的贝叶斯模型对 p_1、p_2 和 p_3 给出的点估计如下

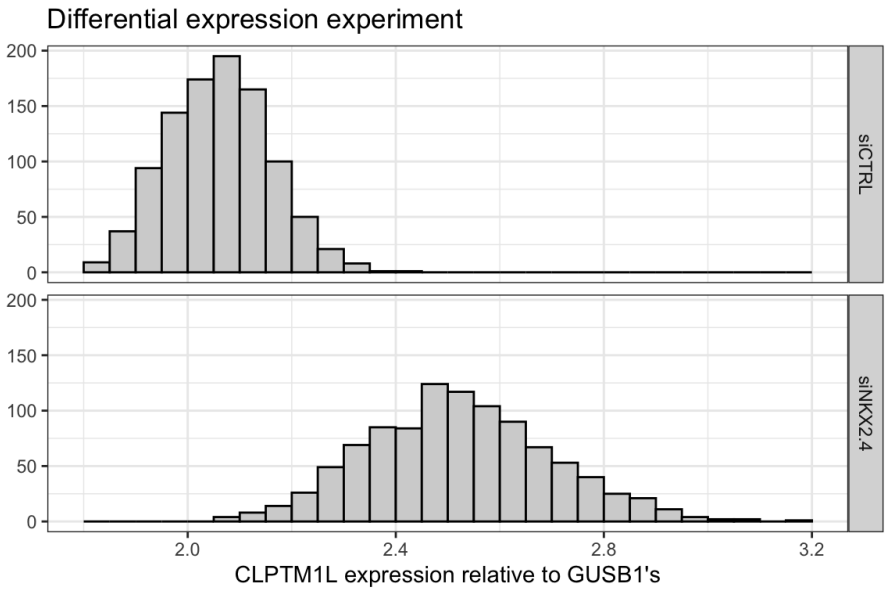


图 5：差异基因表达数据示例。

0.531、0.546 和 0.512，并给出了总体比率的以下可信区间（0.495，0.563）。相比之下，lme4 软件包无法提供稳健的估计，并报告了“边界拟合”和“不良样条拟合”：其 95% 置信区间（0.503，0.547）基于线性插值；鉴于其上限几乎等于一个样本估计值（名称 p_2），该区间似乎极窄。blme 软件包虽能收敛，但最终出现错误消息“Error in zetafun(np, ns) : profiling detected new, lower deviance”

5 讨论

该模型所依据的假设很少，同时足够灵活，可适用于多种情形。实际上，只需提供一个安全的上界 M，用于近似式（2b）和（2c）的连续分布。附录 A 对模型的这一特点作了进一步讨论，但研究人员须注意，在功能基因组学实验中，表达数据可能会给模型带来更大挑战，因为倍数变化很容易达到更极端的数值。尽管如此，默认值（$M = 5000$）旨在为极端倍数变化情形留出充足空间；例如，它似乎足以处理表 2 中所述、在第 3.2 节详细介绍的罕见肿瘤扩增子的检测。事后绘制 α 和 β 的后验分布，确认了这一点：二者均远低于 M 上界。

当 (1) 等位基因比率极大或极小，且同时满足 (2) 大量液滴呈阳性时，二项近似（即假设每个阳性液滴中只有一个扩增子）可能不再成立，估计结果也可能显著偏离真实值。请注意，只有这两个条件同时满足，模型才会失效；例如，分析$GNAQ$ 突变数据（表 2）得到的倍数变化可能很极端，但二项近似仍然成立。我们分析了 bayesddpcr 程序库附带的合成数字滴度 PCR 数据，该数据包含 50 个重复样本，由假设等位基因比率恰为 6 的条件生成，且相当大比例的液滴（总体为 32%）呈阳性。结果高估了真实值，95%可信区间为 (6.9，7.6)；将双阳性液滴计入 A 阳性液滴和 B 阳性液滴的计数后，这次模型低估了倍数变化，且 95%可信区间相当窄，为 (4.9，5.2)。请注意，更复杂的模型（见附录 B）能够得到更令人满意的估计值 (5.6，6.2)（计算成本约为

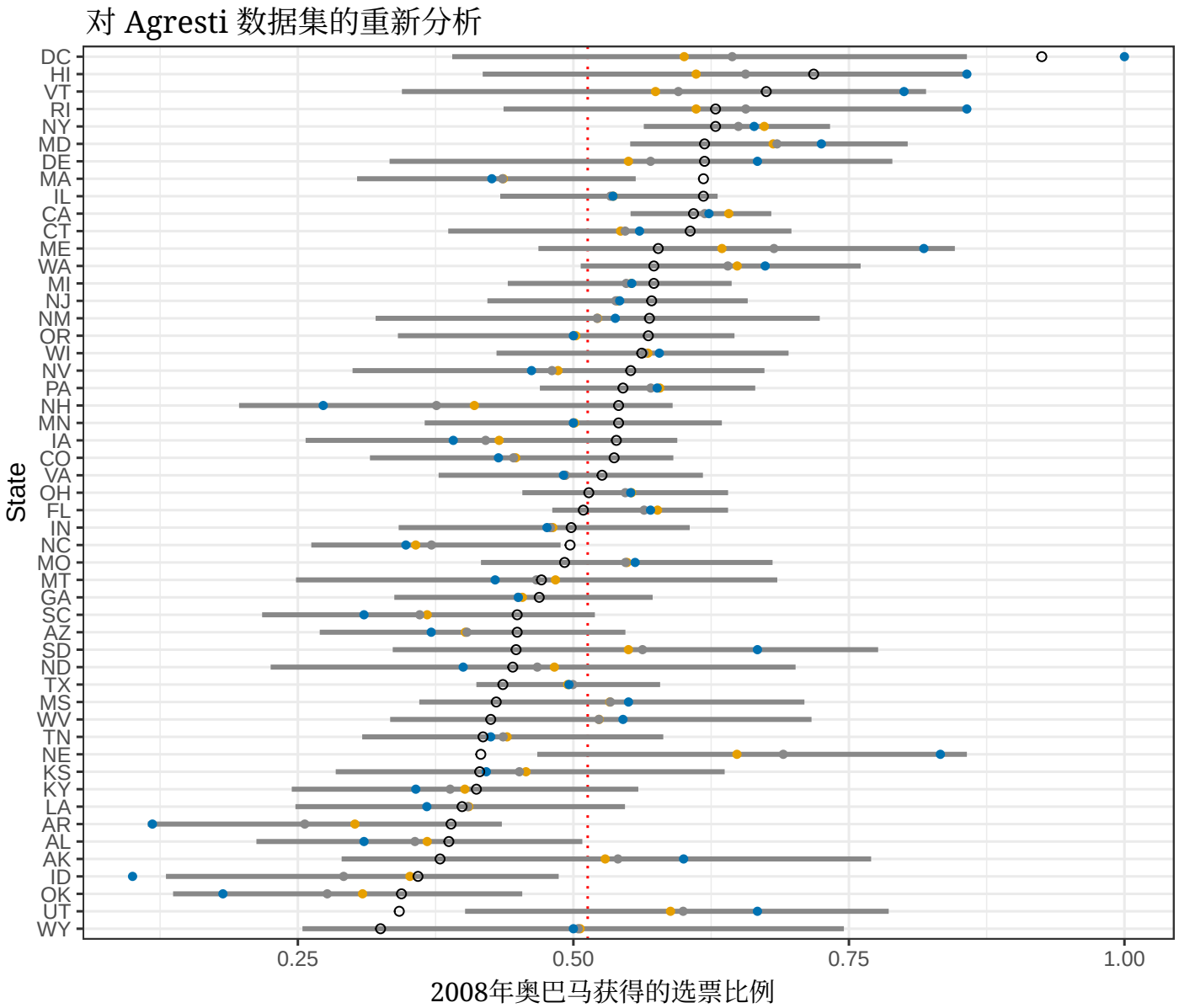


图 6：对 Agresti 数据集的重新分析。灰色圆点和灰色条：州级估计值的中位数和 95%可信区间。橙色圆点：随机效应估计值。蓝色圆点：各样本的比例。黑色圆圈：真实值。Read：全国范围的点估计。

运行时间的十倍）。

在测量差异表达时，假设细胞系彼此独立，可以避免不必要的复杂问题，而且在大多数情况下看起来是合理的，不过这意味着无法据此考虑批次效应。

6 代码可用性

代码以 R 程序包的形式发布于作者的 GitHub 仓库 `https://www.github.com/JosselinNoirel/bayesddpcr/`（可通过 devtools 程序包安装）；用于生成图表的代码和数据集均位于该仓库的“article”目录中。

7 致谢与作者贡献

我们感谢 Anne-Céline Derrien、CarolineHego、Alexandre Houy、Aurore Rampanou Shufang Renault 和 Marc-Henri Stern 提供测试案例，并参与富有启发性的讨论。

Elyas Mouhou 开发了软件、进行了分析并撰写了文章；Vincent Audigier 改进了模型和文章；Josselin Noirel 构思了这项研究、设计了模型并撰写了文章。

A Upp通过众数近似得到的上界

为简便起见，M 上界由人工设定。不过，需要指出的是，可以对式 (2b) 中出现的、由 β、k 和 P 参数化的概率密度函数的众数进行相当精确的近似，如下所示（C 归一化因子）：

$$P(\alpha;\beta,k,P) \propto \mathrm{B}(\alpha,\beta)^{-k} \times P^{\alpha} \quad \alpha > 1 \tag{5}$$

利用 Γ 函数的斯特林近似（$\log\ \Gamma\, z \approx z \log\ z - z$），可以确定众数对应的值 α^* 近似为

$$\alpha^* \approx \beta \cdot \frac{\sqrt[k]{p}}{\sqrt[k]{p}-1}$$

该值可用作初始值，通过迭代求取上界 .

B 基于泊松分布的模型设定

此处给出的二项模型可以进一步改进，以考虑单个液滴中存在多个扩增子的情况；用$\mathcal{N}_i$ 表示第 i 次重复的液滴总数：

$$\alpha \sim \text{均匀分布}(1,\ +\infty) \qquad \text{同上} \tag{6a}$$
$$\beta \sim \text{均匀分布}(1,\ +\infty) \tag{6b}$$
$$p_i \sim \text{贝塔分布}(\alpha,\ \beta) \tag{6c}$$
$$f_i \sim \text{均匀分布}(0,1) \qquad \text{稀释因子} \tag{6d}$$
$$\mathcal{A}_i \sim \text{泊松分布}(f_i \cdot p_i) \qquad \text{A扩增子数量} \tag{6e}$$
$$\mathcal{B}_i \sim \text{泊松分布}(f_i \cdot q_i) \qquad \text{B扩增子数量}（q_i = 1 - p_i） \tag{6f}$$
$$\mathfrak{n}_i = \Pr(\mathcal{A}_i = 0 \wedge \mathcal{B}_i = 0) \qquad \text{阴性样本} \tag{6g}$$
$$\mathfrak{a}_i = \Pr(\mathcal{A}_i \geq 1 \wedge \mathcal{B}_i = 0) \qquad \text{‘A’阳性} \tag{6h}$$
$$\mathfrak{b}_i = \Pr(\mathcal{A}_i = 0 \wedge \mathcal{B}_i \geq 1) \qquad \text{‘B’阳性} \tag{6i}$$
$$\mathfrak{d}_i = \Pr(\mathcal{A}_i \geq 1 \wedge \mathcal{B}_i \geq 1) \qquad \text{双阳性} \tag{6j}$$
$$N_i,\ A_i,\ B_i,\ D_i \sim \text{多项分布}(\mathcal{N}_i,\ \{\mathfrak{n}_i,\ \mathfrak{a}_i,\ \mathfrak{b}_i,\ \mathfrak{d}_i\}) \tag{6k}$$

请注意，尽管吉布斯采样器的计算成本较高，但 bayesddpcr库中已经实现了该采样器。

参考文献

[1] Alan Agresti。《分类数据分析》。Wiley 概率与统计学系列。Wiley- Blackwell，新泽西州霍博肯，第3版，2012年11月。[2] Susana Olmedillas López、Mariano García Arranz 和 Damián García Olmo。微滴数字 PCR 在肿瘤学领域当前及新兴的应用。《分子诊断与治疗》，21(5):493–510，2017年5月。[3] Saw Yen Ow、Malinda Salim、Josselin Noirel、Caroline Evans、Ishtiaq Rehman 和 Phillip C. Wright。简单和复杂混合物中的 iTRAQ 低估：“好、坏与丑”。《蛋白质组学研究杂志》，8(11):5347–5355，2009年9月。[4] John H Schwacke、Elizabeth G Hill、Edward L Krug、Susana Comte-Walters 和 Kevin L Schey。iQuantitator：一种使用 iTRAQ 推断蛋白质表达的工具。《*BMC* 生物信息学》，10(1)，2009年10月。