arXiv:2602.03218v1 [stat.ME] 3 Feb 2026

考虑试验中期估计不确定性的盲态样本量重新估计

Hirotada Maeda

大阪大学医学研究生院生物医学统计学系；国立脑与心血管中心数据科学系，日本大阪府吹田市山田丘 2-2，邮编 565-0871maeda hirotada@biostat.med.osaka-u.ac.jp

Satoshi Hattori

生物医学统计学系、医学研究生院及医学综合前沿研究部门,

开放与跨学科研究倡议研究所（OTRI），大阪大学，日本大阪府吹田市山田丘 2-2, 邮编 565-0871hattoris@biostat.med.osaka-u.ac.jp

Tim Friede

医学统计学系，哥廷根大学医学中心及 DZHK（德国心血管研究中心）和 DZKJ（德国儿童与青少年健康中心），Humboldtallee 32, 37073 哥廷根，德国tim.friede@med.uni-goettingen.de

摘要

为了使随机对照试验能够得出确切结论，在设计阶段准确设定目标样本量至关重要。比较两个正态总体时，

样本量计算需要指定治疗效应以外的方差，而设定不准确可能导致研究检验效能不足。盲态样本量重新估计是一种降低研究无法得出确切结论风险的方法。现有方法提出使用可根据盲态数据估计、且无需了解治疗分配情况的总方差（单样本方差）。我们证明，由于该估计量的期望大于或等于真实方差，因此在盲态审查中，单样本方差方法可视为提供了方差的上界。这种最坏情况评估可能降低研究检验效能不足的风险。然而，对于样本量较小的研究，盲态审查仍可能导致检验效能不足。我们提出一种改进方法，在盲态审查中使用方差的置信上限来考虑估计误差。类似思路此前已用于外部预试验。此外，我们还开发了一种方法，用于选择适当的置信水平，使再估计样本量达到目标检验效能。数值研究表明，我们的方法效果良好，且优于现有方法。所提出的方法以近期随机临床试验为背景进行阐述和说明。

关键词 临床试验 · 内部预试验 · 罕见病 · 样本量重新估计 · 置信上限

1 引言

开展临床试验时，应审慎确定目标样本量，以确保科学性和伦理有效性。确定目标样本量的一种常见方法，是确保对目标治疗效应达到目标检验效能。假设我们正在开展一项临床试验，以比较两种治疗在连续型结局上的效果。由于采用双样本t检验进行最终分析时的检验效能取决于两组正态响应的共同组别特异性方差，因此，准确指定方差对于准确计算样本量至关重要。如果已有一些设计相似的临床试验，可以根据这些研究的结果估计方差（Schmidli等人，2017）。另一方面，通常可获得的信息很少，甚至没有；即使有一些信息，也未必适用于目标研究。因此，在设计阶段很难正确设定方差，而方差设定不当可能导致最终分析的检验效能不足。

据此，可以根据这些研究的结果估计方差（Schmidli等人，2017）。另一方面，通常可获得的信息很少，甚至没有；即使有一些信息，也未必适用于目标研究。因此，在设计阶段很难正确设定方差，而方差设定不当可能导致最终分析的检验效能不足。可以开展一些外部预试验，以获取有关方差的信息。然而，这种做法成本高昂，而且尤其耗时。

为解决这一问题，已有研究提出了基于试验中期数据审查（也称为内部预试验）的样本量重新估计方法（Stein，1945）；此类方法在新药申请中发挥着重要作用（美国食品药品监督管理局（FDA），2019）。基于方差估计的样本量重新估计方法分为两类：非盲态和盲态。在非盲态情境下，内部预试验期间会揭示治疗分配情况，随后可估计两组的治疗效应，并使用双样本合并方差估计量估计方差。Wittes等人（1999）指出，无论设计阶段的方差设定是否正确，非盲态样本量重新估计都能确保检验效能充足。不过，他们也报告称：（i）非盲态样本量重新计算会导致I类错误率膨胀；（ii）若内部预试验纳入的受试者数量较少（例如每组10人），样本量重新计算仍可能导致研究检验效能不足。针对问题（i），已有多种方法被提出用于控制I类错误率（Denne 和 Jennison，1999；Kieser 和 Friede，2000；Miller，2005）。针对问题（ii），Zucker 等人（1999）提出了膨胀因子法，通过用t分布的分位数替代正态分位数，以经验方式增加再估计样本量。尽管方法学方面取得了这些进展，非盲态样本量重新估计仍要求在内部预试验期间获取治疗分配信息，这可能引发对操作偏倚的担忧。为尽量减少操作偏倚，最好在内部预试验期间对分配信息保密（美国食品药品监督管理局（FDA），2019）。因此，为维护临床研究的完整性，盲态审查更为可取。在盲态审查下，可用于样本量重新计算的信息受到限制；无法估计各组的治疗效应，也就无法估计方差。Gould 和 Shih（1992）

提出使用单样本方差，即利用盲态数据可估计的合并总体总方差。单样本方差对共同组别特异性方差存在正偏倚（Zucker 等人，1999），因此可视为方差的上界。Zucker 等人（1999）提出了一种偏倚调整的单样本方差估计量，其中采用假定的目标治疗效应。在优效性试验中，Kieser 和 Friede（2003）表明，与非盲态样本量重新计算不同，使用这些估计量进行盲态样本量重新计算不会导致I类错误率膨胀。此外，Kieser 和 Friede（2003）比较了采用单样本方差和偏倚调整的单样本方差估计量的盲态样本量重新计算方法，并得出结论：未调整的单样本方差估计量更有利于达到目标检验效能。换言之，在单样本方差基础上留出一定余量，有助于避免研究检验效能不足。另一方面，Kieser 和 Friede（2003）报告称，采用单样本方差的小规模内部预试验仍可能导致研究检验效能不足。这表明，单样本方差所提供的余量可能不够，若能更正式地控制这一余量，将会有所裨益。基于这一背景，Friede 和 Kieser（2013）以最初为非盲态样本量重新计算提出的膨胀因子法（Zucker 等人，1999）为基础，考察了该方法在盲态样本量重新估计中的表现，并得出结论：膨胀因子法同样有助于避免盲态情境下研究检验效能不足。然而，该膨胀因子的构造较为经验化，其理论基础尚不明确。

我们提出了一种新的盲态样本量重新估计方法，能够妥善考虑内部预试验中的估计误差。该方法的思路是：通过采用方差的置信上限，并恰当控制置信水平，控制内部预试验中方差估计的误差范围。Browne（1995）将这一思路用于设计阶段的样本量计算，并利用外部预试验的信息。Browne（1995）提出，使用方差的 $100 \cdot (1-\gamma)\%$ 置信上限来确定样本量。基于 Browne（1995）的模拟结果，Kieser and Wassmer（1996）研究了这一方法的理论性质，并推导出一种选择相关

$1-\gamma$的方法，该方法根据外部预试验中纳入的受试者人数来确定，以控制目标检验效能。本文将这一思路扩展到盲态内部预试验情境。将该方法扩展到我们的情境面临两个难点。首先，在盲态审查下，无法一致地估计两组均值，因而也无法一致地估计方差，因此难以明确如何利用盲态数据构造易于处理的方差 $100 \cdot (1-\gamma)\%$ 置信上限。其次，在内部预试验中，用于样本量再计算的数据也会纳入最终分析，因此最终样本量取决于最终分析所用数据的一部分。为解决这些难点，我们借助渐近理论，采用可在盲态审查中估计的保守方差 $100 \cdot (1-\gamma)\%$ 置信上限。为选择能够使最终分析达到目标检验效能的 $1-\gamma$，我们根据给定再估计样本量时的渐近条件功效，推导出功效下界（Zucker 等人，1999）。我们的方法有一个有趣的特点：给定内部预试验的受试者人数后，$1-\gamma$ 可以独立于方差确定。因此，$1-\gamma$ 可以在研究方案中预先指定。通过数值研究，我们发现，即使内部预试验的样本量极小，我们的方法仍能选择合适的 $1-\gamma$，从而避免最终分析的检验效能不足。

我们的工作受到临床场景中难以开展研究的小规模研究的启发。Ban- none 等人（2023）开展了一项针对胰腺切除术后急性胰腺炎的随机临床试验，以评估弥散加权磁共振成像诊断能否快速完成。他们对12名受试者进行了盲态样本量重新计算。Duffley 等（2023）开展了一项针对帕金森病的随机临床试验，以研究帕金森病脑深部电刺激移动应用程序能否缩短帕金森病治疗的程控时间。他们对22名受试者进行了盲态样本量重新计算。这些临床试验表明，一些内部预试验采用的样本量非常小；因此，需要适用于小规模内部预试验的样本量重新估计方法。我们所提出的方法为此提供了一种相关工具。

本文其余部分安排如下。第2节介绍我们假设的设定，并回顾连续型结局的样本量重新计算现有方法。第3节介绍我们所提出的盲态样本量重新计算方法，该方法使用共同组别特异性方差的保守上置信限，并阐述其理论依据。第4节展示数值研究及其结果，以评估所提出方法的性能。第5节通过临床试验的两个真实场景阐释我们所提出的方法，并在第6节进行讨论。

2 现有样本量重新估计方法

2.1 设置

假设我们要设计一项随机试验，比较两种治疗（试验组和对照组）在连续型结局方面的差异。令 $Z \in \{1, 0\}$ 为治疗指示变量， $P(Z = 1) = \pi\,(0 < \pi < 1)$；若受试者被分配至试验组，则 $Z = 1$；若分配至对照组，则 $Z = 0$。令 Y_z 为接受 $N(\mu_z,\ \sigma^2)$ 后的结局，其中 $Z = z$。令 $\Delta := \mu_1 - \mu_0$ 为真实治疗效应， δ 为目标治疗效应，该效应可定义为最小临床重要差异。我们设定目标样本量，使其针对最小临床重要差异的目标检验效能为 $1-\beta$（例如 0.80） δ。考虑以下假设检验问题；

$$H_0 : \Delta = 0 \quad vs. \quad H_1 : \Delta = \delta\,(> 0)$$

我们采用单侧显著性水平为 α 的 Student t检验。然后，可使用以下近似方法计算每组的目标样本量；

$$n_{z,design} = P(Z = z)\Big(\frac{1}{\pi} + \frac{1}{1-\pi}\Big)\Big(\frac{z_\alpha - z_{1-\beta}}{\delta}\Big)^2\sigma^2,$$

其中， z_α 是标准正态分布的上侧 $100 \cdot \alpha$%分位数（Chow等人，2017）。要确定样本量，除了最小临床意义治疗效应δ外，我们还需要指定 σ^2。本文将治疗效应以外的参数称为干扰参数。即使试验治疗实际上具有预期的治疗效应，在设计阶段不恰当地指定 σ^2 也可能导致研究检验效能不足。为解决

这一问题，在内部预试验中重新计算样本量很有用。本文采用盲态样本量重新估计方法，以确保达到目标检验效能。

2.2 正态结局且采用内部预试验时盲态样本量重新估计方法简述

令 $n_{z,int}$ 为纳入内部预试验的受试者人数，$\hat{}n_{z,fin}$ 为治疗组 $Z=z$ 的再估计样本量。两组对应的数值分别记为 $n_{int}=n_{1,int}+n_{0,int}$ 和 $\hat{}n_{fin}=\hat{}n_{1,fin}+\hat{}n_{0,fin}$。组 $Z=z$ 中第 i 名受试者的结局记为 $Y_{z,i}$。不失一般性，我们假设前 $n_{z,int}$ 名受试者纳入内部预试验。在无法获取治疗分配信息的盲态审查中，$\mu_z(z\in\{1,0\})$，因此无法计算两样本合并方差。Gould 和 Shih（1992）提出使用单样本方差；

$$\hat{\sigma}^2_{OS}=\frac{1}{n_{int}-1}\sum_{z=0}^{1}\sum_{i=1}^{n_{z,int}}(Y_{z,i}-\bar{Y}_{int})^2 \tag{2.1}$$

其中，$\bar{Y}_{int}=\sum_{z=0}^{1}\sum_{i=1}^{n_{z,int}}Y_{z,i}/n_{int}$ 是内部预试验中两组的样本均值。Zucker 等人（1999）表明，$\hat{}\sigma^2_{OS}$ 的期望为

$$E[\hat{\sigma}^2_{OS}]=\sigma^2+\frac{n_{1,int}n_{0,int}}{n_{int}(n_{int}-1)}\Delta^2 \tag{2.2}$$

于是，$E[\hat{}\sigma^2_{OS}]\geq\sigma^2$。也就是说，$\hat{}\sigma^2_{OS}$ 是 σ^2 上界的估计值，因此$\hat{}\sigma^2_{OS}$ 很可能会高估 σ^2。通过设定 $\Delta=\delta$，Zucker 等人（1999）提出了调整后单样本方差，以校正单样本方差的偏差项；

$$\hat{\sigma}^2_{ADJ}=\hat{\sigma}^2_{OS}-\frac{n_{1,int}n_{0,int}}{n_{int}(n_{int}-1)}\delta^2$$

根据（2.2），如果 $\Delta=\delta$ 成立，则$\hat{}\sigma^2_{ADJ}$ 是 σ^2的无偏估计量；$E[\hat{}\sigma^2_{ADJ}]=\sigma^2$。通过用$\hat{}\sigma^2=\hat{}\sigma^2_{OS}$ 或$\hat{}\sigma^2=\hat{}\sigma^2_{ADJ}$估计 σ^2，组 $Z=z$ 的样本量将重新估计为

$$\hat{n}_{z,fin}=P(Z=z)\Big(\frac{1}{\pi}+\frac{1}{1-\pi}\Big)\Big(\frac{z_\alpha-z_{1-\beta}}{\delta}\Big)^2\hat{\sigma}^2 \tag{2.3}$$

然后，重新计算的总样本量为

$$\hat{n}_{fin}=\Big(\frac{1}{\pi}+\frac{1}{1-\pi}\Big)\Big(\frac{z_\alpha-z_{1-\beta}}{\delta}\Big)^2\hat{\sigma}^2 \tag{2.4}$$

样本量重新估计后，在观察到$\hat{n}_{fin}$名受试者的所有结局后进行最终分析。检验统计量 T 为

$$T = \frac{\bar{Y}_{1,fin} - \bar{Y}_{0,fin}}{\sqrt{\hat{\sigma}^2_{fin}(1/\hat{n}_{1,fin} + 1/\hat{n}_{0,fin})}}$$

其中，$\bar{Y}_{z,fin} = \sum_{i=1}^{\hat{n}_{z,fin}} Y_{z,i}/\hat{n}_{z,fin}$ 和 $\hat{\sigma}^2_{fin} = \sum_{z=0}^{1}\sum_{i=1}^{\hat{n}_{z,fin}} (Y_{z,i} - \bar{Y}_{z,fin})^2/(\hat{n}_{fin} - 2)$分别表示最终分析时各组的样本均值和双样本合并方差估计量。

Kieser 和 Friede（2003）评估了基于（2.4）的样本量重新计算的检验效能，其中使用 $\hat{\sigma}^2_{OS}$和 $\hat{\sigma}^2_{ADJ}$。考虑给定 $\hat{n}_{fin}$时的条件检验效能 $P(T \geq t_{\hat{n}_{fin}-2,\alpha} | \hat{n}_{fin}, H_1)$，其中$t_{\hat{n}_{fin}-2,\alpha}$ 是自由度为 $\hat{n}_{fin}$ −2 的 t 分布的上侧 $100 \cdot \alpha$% 分位数，通过检验统计量 T的精确（而非渐近）分布计算得到。由于样本量 $\hat{n}_{fin}$ 取决于内部预试验阶段的数据，因此通过对给定 $\hat{n}_{fin}$时的条件检验效能 $P(T \geq t_{\hat{n}_{fin}-2,\alpha} | \hat{n}_{fin}, H_1)$ 关于 $\hat{n}_{fin}$的分布求边际，推导出检验效能。他们得出结论：在特定备择假设下无偏的 $\hat{\sigma}^2_{ADJ}$ 可能导致最终分析检验效能不足，并建议使用 $\hat{\sigma}^2_{OS}$ 以避免研究结果不确定。然而，正如 Kieser 和 Friede（2003）的表2所示，当 n_{int} 较小时，采用 $\hat{\sigma}^2_{OS}$ 的样本量重新估计方法不一定能维持目标检验效能。为解决这一问题，Friede 和 Kieser（2013）将膨胀因子法（Zucker 等人，1999）应用于盲态样本量重新估计，并考察了将再估计样本量（2.4）乘以膨胀因子 $IF = (t_{n_{int}-2,\alpha} - t_{n_{int}-2,1-\beta})^2/(z_\alpha - z_{1-\beta})^2$, 后所得样本量下的检验效能，其中 $\hat{\sigma}^2 = \hat{\sigma}^2_{OS}$;

$$\hat{n}_{fin} = \left(\frac{1}{\pi} + \frac{1}{1-\pi}\right)\left(\frac{t_{n_{int}-2,\alpha} - t_{n_{int}-2,1-\beta}}{\delta}\right)^2 \hat{\sigma}^2_{OS} \tag{2.5}$$

3 考虑试验中期估计不确定性的盲态样本量重新估计及置信上限

3.1 所提方法的动机与思路

如第2.2节所述，使用 $\hat{\sigma}^2_{OS}$ 被视为在基于内部预试验盲态数据、以上界为限的 σ^2估计中留出一定余量。然而，为什么应使用这一上界尚不清楚；更恰当地处理高估问题，可能会使该方法在检验效能和样本量分布方面具有更好的性质。

在本文中，我们考虑在内部预试验中利用置信区间来控制检验效能。也就是说，我们考虑在盲态审查下使用 $\hat{\sigma}^2_{OS}$ 构造方差的置信上限来重新估计样本量，并通过选择适当的置信水平 $100 \cdot (1-\gamma)$ %，使其达到目标检验效能。在设计阶段的样本量计算背景下，已有一些方法考虑了基于外部预试验估计的置信上限（Browne，1995；Kieser and Wassmer，1996）。Kieser and Wassmer（1996）指出，可以确定 $1-\gamma$，使检验效能等于目标检验效能，因为在给定最终统计分析的显著性水平时，检验效能仅由 $1-\gamma$ 和外部预试验的受试者人数决定。在本研究中，我们将他们的思路应用于内部预试验中的盲态样本量重新估计。需要考虑两点。首先，由于 Kieser and Wassmer（1996）研究的是外部预试验，用于重新估计样本量的外部预试验估计参数与最终统计量在统计上相互独立。其次，Kieser and Wassmer（1996）关注的是非盲情形，因此可以估计组别特异性参数，进而估计 σ^2的置信上限。这就是他们的检验效能公式可以不包含效应量 Δ/σ，并据此确定 $1-\gamma$，使最终分析达到目标检验效能的原因。在盲态样本量重新计算的背景下，Kieser and Friede（2003）利用 t 检验的精确分布（而非渐近分布）计算其检验效能。除 $1-\gamma$ 和 n_{int}外，他们的检验效能还取决于效应量 Δ/σ。请注意，效应量未知，因此 Kieser and Friede（2003）仅将检验效能公式用于评估样本量重新计算方法的性能，而未将其用于确定样本量。此外，使用 $\hat{\sigma}^2_{OS}$ 构造的 σ^2 置信上限也取决于非中心卡方分布的非中心参数 λ，而该参数又取决于效应量 Δ/σ （见第3.2节）。因此，在盲态审查中无法估计该置信上限，要应用 Kieser and Wassmer（1996）的思路，使用 $100 \cdot (1-\gamma)$ %置信上限来计算检验效能，是一项颇具挑战的任务。为克服这些困难，需要一种新思路；我们建议使用保守上置信限，

可以在盲态审查下构建，并推导出不依赖于效应量 Δ/σ 的检验效能下界。通过设定 $1-\gamma$，使检验效能下界达到目标检验效能，可以避免研究检验效能不足。

3.2 渐近有效但不可行的置信区间上限方法

首先，我们说明如何利用内部预试验构建关于 σ^2 的渐近有效置信区间，以及如何根据置信区间上限计算检验效能的公式。该公式被证明依赖于 Δ/σ，因此无法应用于盲态审查。尽管实际上不可行，我们仍在此考虑该方法，以阐明我们思路的某些方面。设 W 为服从 $\chi^2_{n_{int}-1}$ 的随机变量，即自由度为 $n_{int}-1$ 的中心卡方分布；设 W_λ 为服从 $\chi^2_{n_{int}-1,\ \lambda}$ 的随机变量，即自由度为 $n_{int}-1$、非中心参数为 λ 的非中心卡方分布。分别用 F_W、f_W 和 $d_{1-\gamma}$ 表示 $\chi^2_{n_{int}-1}$ 的累积分布函数、概率密度函数和上侧 $100\cdot\alpha\%$ 分位数。非中心卡方分布 $\chi^2_{n_{int-1,\ \lambda}}$ 的相应函数和分位数分别记为 F_{W_λ}、f_{W_λ} 和 $d_{\lambda,\ 1-\gamma}$。注意，$\hat{}\sigma^2_{OS}(n_{int}-1)/\sigma^2 \sim \chi^2_{n_{int}-1,\ \lambda}$ 的非中心参数为 $\lambda=(\Delta/\sigma)^2(n_{1,\ int}n_{0,\ int})/n_{int}$。于是，有

$$P\Big(\frac{\hat{\sigma}^2_{OS}(n_{int}-1)}{d_{\lambda,1-\gamma}}\geq\sigma^2\Big)=P\Big(\frac{\hat{\sigma}^2_{OS}(n_{int}-1)}{\sigma^2}\geq d_{\lambda,1-\gamma}\Big)=P(W_\lambda\geq d_{\lambda,1-\gamma})=1-\gamma.$$

因此，σ^2 的 $100\cdot(1-\gamma)\%$ 置信上限为

$$\hat{\sigma}^2_{U,1-\gamma}=\hat{\sigma}^2_{OS}\frac{n_{int}-1}{d_{\lambda,1-\gamma}} \tag{3.1}$$

由于 $\hat{}\sigma^2_{U,\ 1-\gamma}$ 依赖于 $d_{\lambda,\ 1-\gamma}$，而非中心参数 λ取决于未知效应量 Δ/σ，因此在盲态审查下无法估计。我们给出基于两样本方差估计量 $\hat{}\sigma^2_{fin}$ 的检验效能公式如下。该公式在盲态下不可行，但对我们的推导有用。由于检验统计量 T 中包含 $\hat{}\sigma^2_{fin}$，且 $t_{\hat{n}_{fin}-2,\ \alpha}$ 分别渐近近似于 σ^2 和 z_α，因此给定 $\hat{}n_{fin}$ 时的渐近条件功效表示为

如下（Zucker 等人，1999）；

$$P(T \geq t_{\hat{n}_{fin}-2,\alpha}|\hat{n}_{fin}, H_1) = P\Big(\frac{\bar{Y}_{1,fin} - \bar{Y}_{0,fin}}{\sqrt{\hat{\sigma}^2_{fin}(1/\hat{n}_{1,fin} + 1/\hat{n}_{0,fin})}} \geq t_{\hat{n}_{fin}-2,\alpha}|\hat{n}_{fin}, H_1\Big)$$
$$\approx P\Big(\frac{\bar{Y}_{1,fin} - \bar{Y}_{0,fin}}{\sqrt{\sigma^2(1/\hat{n}_{1,fin} + 1/\hat{n}_{0,fin})}} \geq z_\alpha|\hat{n}_{fin}, H_1\Big)$$
$$= 1 - \Phi\Big(z_\alpha - \frac{\Delta}{\sqrt{\sigma^2\{1/\pi + 1/(1-\pi)\}/\hat{n}_{fin}}}\Big) \tag{3.2}$$

根据 (2.3) 计算 $\hat{n}_{fin}$ with $\hat{\sigma}^2_{U,\ 1-\gamma}$；

$$\hat{n}_{fin} = \Big(\frac{1}{\pi} + \frac{1}{1-\pi}\Big)\Big(\frac{z_\alpha - z_{1-\beta}}{\delta}\Big)^2 \hat{\sigma}^2_{U,1-\gamma} \tag{3.3}$$

将 (3.3) 代入 (3.2) 中的 $\hat{n}_{fin}$，得到

$$P(T \geq t_{\hat{n}_{fin}-2,\ \alpha}|\hat{n}_{fin},\quad H_1)$$
$$\approx 1 - \Phi\Big(z_\alpha - \frac{\Delta}{\delta} \cdot |z_\alpha - z_{1-\beta}| \cdot \sqrt{\frac{\hat{\sigma}^2_{U,1-\gamma}}{\sigma^2}}\Big) \tag{3.4}$$

由于 $\hat{\sigma}^2_{U,\ 1-\gamma}/\sigma^2 = W_\lambda/d_{\lambda,\ 1-\gamma}$ 以及 (3.4)，检验效能表示为

$$P(T \geq t_{\hat{n}_{fin}-2,\alpha}|H_1) \approx 1 - \int_0^\infty \Phi\Big(z_\alpha - \frac{\Delta}{\delta} \cdot |z_\alpha - z_{1-\beta}| \cdot \sqrt{\frac{w}{d_{\lambda,1-\gamma}}}\Big) f_{W_\lambda}(w)dw \tag{3.5}$$

(3.5) 的右侧由 $d_{\lambda,\ 1-\gamma}$ 和 W_λ 表示。由于 λ包含未知的效应量 Δ/σ，因此在实践中无法计算 (3.1) 和 (3.5)。

3.3 所提出的方法：在 σ^2 盲态审查中使用可估计量的保守上置信限

符号汇总于表 1 中。以下我们将使用具有相同自由度的中心卡方分布与非中心卡方分布之间关系的一些性质。首先给出以下引理。

引理 1 对于具有相同自由度的中心卡方分布和非中心卡方分布，有以下不等式成立：

$$F_{W_\lambda}(w) \leq F_W(w)\ (\forall w \geq 0)$$

因此，下列不等式也成立：

$$d_{\lambda,1-\gamma} \geq d_{1-\gamma}$$

引理2 设 F_1, F_2 为 X_1, X_2的分布函数，满足 $F_1(x) \le F_2(x)\ (\forall x \in \mathbb{R})$。假设非递减函数 g 满足 $E[g(X_1)] < \infty$且 $E[g(X_2)] < \infty$，则下列不等式成立：

$$E[g(X_1)] \ge E[g(X_2)]$$

引理 1 和引理 2 的证明分别见 Finner and Roters（1997）和 Lehmann（1955，第 757–777 页）。定义 $\bar{\sigma}^2_{U,\ 1-\gamma} = \hat{\sigma}^2_{U,\ 1-\gamma} d_{\lambda,\ 1-\gamma}/d_{1-\gamma}$。由于（3.1），

$$\bar{\sigma}^2_{U,1-\gamma} = \hat{\sigma}^2_{U,1-\gamma}\frac{d_{\lambda,1-\gamma}}{d_{1-\gamma}} = \hat{\sigma}^2_{OS}\frac{n_{int}-1}{d_{1-\gamma}} \tag{3.6}$$

成立。注意，它不依赖于 Δ/σ。由（3.6）可得 $\bar{\sigma}^2_{U,\ 1-\gamma} d_{1-\gamma} = \hat{\sigma}^2_{OS}(n_{int}-1)$，于是 $\bar{\sigma}^2_{U,\ 1-\gamma} d_{1-\gamma}/\sigma^2$ 服从非中心卡方分布 W_λ，因为 $\hat{\sigma}^2_{OS}(n_{int}-1)/\sigma^2 \sim W_\lambda$。也就是说，$\bar{\sigma}^2_{U,\ 1-\gamma}/\sigma^2 = W_\lambda/d_{1-\gamma}$ 成立。这是我们推导中的一个关键量。可以得到

$$P(\bar{\sigma}^2_{U,1-\gamma} \ge \sigma^2) = P\Big(\hat{\sigma}^2_{U,1-\gamma}\cdot\frac{d_{\lambda,1-\gamma}}{d_{1-\gamma}} \ge \sigma^2\Big) \ge P(\hat{\sigma}^2_{U,1-\gamma} \ge \sigma^2) = 1-\gamma$$

该不等式由引理 1 得出。因此，$\bar{\sigma}^2_{U,\ 1-\gamma}$ 为我们提供了一个保守的 $100\cdot(1-\gamma)\%$ 置信上限，用于 σ^2。根据（2.3），令 $\bar{\sigma}^2_{U,\ 1-\gamma}$，计算 $\hat{n}_{fin}$；

$$\hat{n}_{fin} = \Big(\frac{1}{\pi}+\frac{1}{1-\pi}\Big)\Big(\frac{z_\alpha - z_{1-\beta}}{\delta}\Big)^2 \bar{\sigma}^2_{U,1-\gamma} \tag{3.7}$$

将（3.7）代入（3.2）中的 $\hat{n}_{fin}$，得

$$\begin{aligned} &P(T \ge t_{\hat{n}_{fin}-2,\ \alpha} | \hat{n}_{fin},\ \ H_1) \\ &\approx 1-\Phi\Big(z_\alpha - \qquad\qquad \frac{\Delta}{\delta}\cdot|z_\alpha - z_{1-\beta}|\cdot\sqrt{\frac{\bar{\sigma}^2_{U,\ 1-\gamma}}{\sigma^2}}\Big) \end{aligned} \tag{3.8}$$

成立。由 $\bar{\sigma}^2_{U,\ 1-\gamma}/\sigma^2 = W_\lambda/d_{1-\gamma}$ 和（3.8），检验效能为

$$P(T \ge t_{\hat{n}_{fin}-2,\alpha}|H_1) \approx 1-\int_0^\infty \Phi\Big(z_\alpha - \frac{\Delta}{\delta}\cdot|z_\alpha - z_{1-\beta}|\cdot\sqrt{\frac{w}{d_{1-\gamma}}}\Big) f_{W_\lambda}(w)dw \tag{3.9}$$

公式 (3.9) 仍包含非中心参数 $\lambda = (\Delta/\sigma)^2(n_{1,\ int}n_{0,\ int})/n_{int}$，该参数取决于效应量 Δ/σ。接下来，我们考虑 (3.9) 的一个下界。我们将引理 2 应用于 W 和 W_λ的分布。令 $g(w)$ 为 (3.9) 的被积函数；

$$g(w) = 1-\Phi\Big(z_\alpha - \frac{\Delta}{\delta}\cdot|z_\alpha - z_{1-\beta}|\cdot\sqrt{\frac{w}{d_{1-\gamma}}}\Big)$$

这里，在 H_1：$\Delta = \delta$ 条件下， $g(w)$ 是非递减函数。尽管一般而言， $g(w)$ 并非非递减函数，但为计算功效下界，只需考虑 H_1 条件下的 $g(w)$。检验效能 (3.9) 表示为 $E[g(W_\lambda)]$。根据引理 1， $F_{W_\lambda}(w) \leq F_W(w)$ $(\forall w \geq 0)$ 成立。接着，根据引理 2，

$$\begin{aligned} E[g(W_\lambda)] &\geq E[g(W)] \\ &= 1 - \int_0^\infty \Phi\Big(z_\alpha - \frac{\Delta}{\delta} \cdot |z_\alpha - z_{1-\beta}| \cdot \sqrt{\frac{w}{d_{1-\gamma}}}\Big) f_W(w) \end{aligned} \tag{3.10}$$

成立。它给出了（3.9）中的检验效能下界 $E[g(W_\lambda)]$ ，且不依赖于 λ 和 σ^2。在H_1：$\Delta = \delta$的条件下，给定 n_{int} 和最终分析的显著性水平 α ，可以计算仅取决于 $1-\gamma$ 的（3.10），然后确定 $1-\gamma$，使（3.10）达到目标检验效能。利用所得的 $1-\gamma$，可以通过（3.7）确定 $\hat{n}_{fin}$，从而确保达到目标检验效能。总之，我们提出以下步骤;

步骤 1.（设计阶段）设定单侧显著性水平 α、目标检验效能 $1-\beta$、目标治疗效应 δ 以及盲态样本量重新计算的样本量 n_{int}。

步骤 2.（设计阶段）确定满足（3.10）$= 1-\beta$的 $1-\gamma$，并将其写入研究方案。步骤 3.（内部预试验）计算 $\bar{\sigma}^2_{U,\ 1-\gamma}$，然后通过（3.7）确定 $\hat{n}_{fin}$。

步骤4。（最终分析）使用$\hat{n}_{fin}$ 名受试者进行最终分析。

所提出的方法有一个有趣的特点，即 $1-\gamma$ 可以独立于数据确定。表2列出了在 n_{int}的某些设置下的 $1-\gamma$。研究方案中也可以针对 n_{int} 的一些合理取值范围提供类似的表格。

4 数值研究

本节评估所提出方法的性能。我们关注以下问题：(i) 所提出的方法能否良好地控制检验效能，从而有助于避免检验效能不足的研究；(ii) 所提出的盲态样本量重新计算方法是否会导致 I 类错误率膨胀；以及 (iii) 重新计算的样本量服从何种分布。对于 (i) 和 (ii)，我们

评估了 Kieser 和 Friede（2003）给出的 I 类错误率和检验效能的理论值；公式见 Kieser 和 Friede（2003）第6页第11行和第5页第13行。为计算公式中的积分，我们没有采用数值积分，而是使用蒙特卡洛积分，并生成了10^8 均匀随机数。我们首先按照 Kieser 和 Friede（2003）的设定评估检验效能，以探讨问题 (i)，具体设定如下。假设我们正在进行一项以连续型结局为指标的随机临床试验。受试者按分配比例 $\pi = 0.5$随机分配至治疗组或对照组。我们设定 t 检验的目标样本量，单侧显著性水平为 $\alpha = 0.025$，目标检验效能为 $1 - \beta = 0.80$，目标治疗效应为 $\delta = 1$。我们考虑三种方差设定：$\sigma^2 = 2.038$、4.013 和 11.08。要达到目标检验效能，各组分别需要纳入 $N = 32$、63 和 174 名受试者。我们进行了盲态样本量重新计算，并采用以下方法：Gould 和 Shih （1992）在（2.4）中提出的单样本方差方法，称为单样本方法；Zucker 等人（1999）在（2.5）中提出的膨胀因子法，称为 IF；以及使用（3.7）中的方差估计值 $\bar{\sigma}^2_{U、1-\gamma}$ 的所提方法，称为 Proposed。我们还采用了第3.2节中使用（3.3）里的 $\hat{}\sigma^2_{U、1-\gamma}$ 的方法作为参照，称为理论方法。实际操作中，即使分配比例为 $\pi = 0.5$，盲态审查时的样本量也可能不相等。Kieser 和 Friede（2003）在假设 $n_{1、int} = n_{0、int}$ 的条件下计算了检验效能。相比之下，我们没有采用这一假设，而是推导了检验效能并进行了数值研究。不过，由于样本量不等时结果相似，我们仅展示 $n_{1、int} = n_{0、int}$ 的结果。表3展示了检验效能。与 Kieser 和 Friede（2003）的结果一致，当内部预试验的样本量极小时，单样本方法的检验效能可能低于 0.80 的目标检验效能。例如，内部预试验仅有 10 名受试者时（$n_{z、int} = 5$），单样本方法在 $N = 32$ 下的检验效能为 0.7517。另一方面，所提出的方法很好地控制了检验效能。例如，内部预试验有 10 名受试者时，所提方法在 $N = 32$ 下的检验效能为 0.8153。如表3所示，Zucker 等人（1999）的 IF 方法可以避免检验效能不足的研究。然而，与所提方法相比，IF 方法对检验效能的高估幅度更大，

$N = 32$ 和 63。无论研究规模如何，所提出的方法都成功控制了检验效能。有趣的是，理论方法未能达到目标检验效能。请注意，考虑到根据 (3.10) 进行的保守检验效能评估，所提出的方法比理论方法采用了更大的裕量。因此，重新计算的样本量更准确。

接下来，我们评估 I 类错误率。使用 Kieser 和 Friede (2003) 的公式并结合$\Delta = 0$，在表 4 中评估了 $N = 32$、63 和 174 对应的最终分析中使用重新计算的样本量时的 I 类错误率。所有方法的 I 类错误率均未显著膨胀。

最后，我们评估最终分析所用重新计算样本量的变异性。由于 $\hat{}n_{z,\ fin}$ 是 $\hat{}\sigma^2_{OS}$ 的函数（见 (2.4)），且 $\hat{}\sigma^2_{OS}$ 服从非中心卡方分布，因此可以通过计算机生成 $\hat{}\sigma^2_{OS}$，研究单样本方法下重新计算的样本量 $\hat{}n_{z,\ fin}$ （参见 Kieser 和 Friede (2003) 第 5 页第 6 行）。类似地，对于 IF、理论方法和所提出的方法，可分别使用 (2.5)、(3.3) 和 (3.7) 生成重新计算的样本量。表 5 汇总了单样本、IF、所提出的方法和理论方法生成的 10^8 $\hat{}n_{z,\ fin}$ 的均值和经验标准差。具体而言，图1比较了单样本方法和所提出的方法下 $\hat{}n_{z,\ fin}$ 的分布，展示了中位数、 1st 和 3rd 四分位数。结果表明，当内部预试验的样本量极小（例如总计 4 人）时，重新计算的样本量 $\hat{}n_{z,\ fin}$ 变异性很大。如前所述，即使内部预试验只有 4 名受试者，所提出的方法的检验效能仍接近目标检验效能，因此能够很好地控制检验效能。然而，内部预试验的样本量极小时，我们必须付出重新计算的样本量变异性较大的代价。

5 临床试验示例

5.1 一项针对胰腺切除术后急性胰腺炎的随机临床试验

Bannone 等（2023）研究了弥散加权磁共振成像（DW-MRI）诊断对胰腺切除术后急性胰腺炎（PPAP）和术后是否能够迅速起效

高淀粉酶血症（POH）。主要结局为 DW-MRI 上的胰腺表观扩散系数（ADC）值，并将其均值在患有 PPAP 和 POH 的患者与未患 PPAP 或 POH 的患者之间进行比较。Bannone 等（2023）在以下假设下确定了样本量 $n_{design} = 24$； $\alpha = 0.05$ （双侧），$1-\beta = 0.85$， $\Delta = \delta = 4.5\times10^{-4}$， $\sigma^2 = 1.35\times10^{-7}$。Bannone 等（2023）采用单样本方差 $\hat{}\sigma^2_{OS}$ 在 $n_{int} = 12$ 进行了样本量重新估计，样本量被重新估计为 $\hat{}n_{fin} = 65$。在最终分析中，治疗效应被估计为 $\hat{}\Delta = 2.3 \times 10^{-4}$，p 值计算为0.006。

在此设定下，我们采用所提出的方法计算（3.7）。为实施该方法，我们假设结局服从正态分布，且采用均衡分配（$n_{1,\ int} = n_{0,\ int}$， $\pi = 0.5$）。尽管 Bannone 等（2023）未报告 $\hat{}\sigma^2_{OS}$，我们根据其样本量重新估计的结果计算得出 $\hat{}\sigma^2_{OS} = 3.67 \times 10^{-7}$。给定 $n_{int} = 12$，我们计算满足（3.10）的 $1 - \gamma = 0.62$，使其等于目标检验效能 $1 - \beta = 0.85$。最后，在 $\hat{}\sigma^2_{OS} = 3.67 \times 10^{-7}$、$1 - \gamma = 0.62$、 $\alpha = 0.025$ （单侧）和 $1 - \beta = 0.85$ 的设定下，我们得到 $\bar{\sigma}^2_{U,\ 1-\gamma} = \bar{\sigma}^2_{U,\ 0.62} = 4.48 \times 10^{-7}$，随后使用（3.7）计算 $\hat{}n_{fin} = 80$。因此，尽管 Bannone 等（2023）成功得到了具有统计学显著性的结果（$p = 0.006$），但单样本方差 $\hat{}\sigma^2_{OS}$ 可能过小，无法确保达到目标检验效能。这与数值研究中的观察结果一致。

5.2 一项针对帕金森病的随机临床试验

Duffley 等（2023）研究了帕金森病深部脑刺激移动应用程序（MAP DBS）是否能够缩短脑深部电刺激（DBS）的程控时间；脑深部电刺激是治疗帕金森病（PD）的有效方法。作为主要结局指标，研究采用对数转换估算标准治疗脑深部电刺激（SOC DBS）组与 MAP DBS 组之间的平均总程控时间差异。Duffley 等（2023）在以下假设下确定了样本量 $n_{design} = 72$： $\alpha = 0.05$ （双侧）、 $1 - \beta = 0.80$、 $\Delta = \delta = 0.40$， $\sigma^2 = 0.362$。Duffley 等（2023）在 $n_{int} = 22$ 时，使用单样本方差 $\hat{}\sigma^2_{OS} = 0.192$ 进行了样本量重新计算。Duffley 等（2023）的论文未报告重新计算后的样本量；我们根据计算结果将其重建为 $n_{int} = 38$。

最后，Duffley 等（2023）未调整样本量，并使用 $\hat{n}_{design} = 72$ 进行了最终分析，报告治疗效应为 $\Delta = 0.47$，p 值为 0。001。

在此设定下，我们应用所提方法计算（3.7）。为此，我们假设结局服从对数正态分布，并采用均衡分配（$n_{1,int} = n_{0,int}, \pi$ =0.5）。给定 $n_{int} = 22$，我们计算使（3.10）等于目标检验效能$1 - \beta = 0.80$的 $1 - \gamma = 0.57$。最后，在$\hat{\sigma}^2_{OS} = 0.192$、 $1 - \gamma = 0.57$、 $\alpha = 0.025$ （单侧）、$1 - \beta = 0.80$的设定下，我们得到$\bar{\sigma}^2_{U,1-\gamma} = \bar{\sigma}^2_{U,0.57} = 0.210$，然后用（3.7）计算$\hat{n}_{fin} = 42$。因此，如果Duffley 等（2023）将样本量由原来的72改为38而非42，则可能导致最终分析中检验效能不足。

6 讨论

本文提出了一种盲态样本量重新计算方法。由于内部预试验中的估计误差，与检验效能相关的干扰参数的无偏估计量可能导致检验效能不足的研究。相反，在估计中留出一定裕度将改善样本量的表现。与广泛使用的单样本方差方法相比，我们的方法提供了更为规范的留出裕度的方式。该方法成功克服了单样本方差方法的一个缺陷；即使内部预试验的样本量小于20，我们的方法仍能控制检验效能，并且令人惊讶的是，在内部预试验样本量极小（如 $n_{int} = 4$）时仍能维持目标检验效能。

然而，我们需要再次强调，内部预试验的规模极小时，就必须承受重新计算的样本量变异很大的代价，样本量重新计算可能会给出不切实际的巨大样本量建议（例如，结合亚组分析的盲态样本量重新计算；Placzek 和 Friede（2023））。通常，样本量重新计算用于受试者数量确定较大的确证性临床试验，因此通常是在中等或较大的样本量下进行。然而，正如第5节所示，有些临床试验采用规模极小的内部预试验。从避免因研究检验效能不足而导致试验失败的风险这一角度来看，对于难以招募患者的罕见病临床试验而言，这种做法可能很有帮助；不过，重新计算的样本量^{v1}变异性较大的问题应得到适当处理。

效能不足的研究，同时应妥善处理重新计算的样本量 $\hat{n}_{fin}$ 变异性较大的问题。

我们所提方法的成功之处在于引入了对检验效能的保守评估。评估所用的关键技术工具，来自中心卡方分布与非中心卡方分布之间的关系，这些关系概述于第3节的引理 1 和引理 2。尽管这些关系早已确立（Lehmann，1955；Finner and Roters，1997），但据我们所知，它们尚未用于开发自适应设计方法。我们预计，本文采用的技术也有助于解决其他自适应设计问题。

尽管本文重点讨论盲态样本量重新计算，我们的理论也适用于非盲态情境。可以考虑使用基于双样本合并方差估计量构造的置信上限（例如 Kieser and Wassmer（1996）所提供的方法），该估计量可在非盲态审查中估计。Kieser 和 Friede（2000）针对不同的置信水平 $1-\gamma$（范围从 $60-90\%$）考虑了这一程序，并控制了 I类错误率。根据本文提出的程序，可以扩展他们的方法，推导出仅取决于 $1-\gamma$ 的检验效能，并确定 $1-\gamma$，从而达到目标检验效能。利用由此得到的 $1-\gamma$，即使内部预试验的样本量很小，也可以重新估计样本量，以确保达到目标检验效能。

参考文献

Bannone, E., Marchegiani, G., Zamboni, G. A., Maris, B. M., Costa, L., Procida, G., Vacca, P. G., D'Onofrio, M., Mansueto, G., De-Madaria, E., 等（2023）。胰十二指肠切除术后急性胰腺炎：一项关于磁共振扩散加权成像、血清生物标志物和临床特征的前瞻性研究。《Surgery》，173(6):1428–1437。

Browne, R. H.（1995）。关于使用预试样本进行样本量确定。Statistics in Medicine, 14(17):1933–1940。

Chow、S.-C.、Shao、J.、Wang、H. 和 Lokhnygina、Y.（2017）。临床研究中的样本量计算。Chapman and Hall/CRC。

Denne、J. S. 和 Jennison、C.（1999）。使用内部预试验估计t检验的样本量。Statistics in Medicine，18(13)：1575–1585。

Duffley, G., Szabo, A., Lutz, B. J., Mahoney-Rafferty, E. C., Hess, C. W., Ramirez-Zamora, A., Zeilman, P., Foote, K. D., Chiu, S., Pourfar, M. H. 等（2023）。帕金森病脑深部电刺激交互式移动应用程序（MAP DBS）：一项开放标签、多中心、随机对照临床试验。《帕金森病与相关疾病》，109:105346。

Finner, H. 和 Roters, M.（1997）。对数凹性以及卡方、F 和贝塔分布的不等式及其在多重比较中的应用。《中华统计学志》，7(3):771–787。

美国食品药品监督管理局（FDA）（2019）。药物和生物制品适应性设计临床试验：行业指南。https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry。

Friede, T. 和 Kieser, M.（2013）。优效性和非劣效性试验中的盲态样本量重新估计：方差估计中的偏倚与方差。《药物统计学》，12(3)：141–146。

Gould, A. L. 和 Shih, W. J.（1992）。针对正态分布结局且方差未知的情况，在不揭盲的前提下重新估计样本量。《统计通讯：理论与方法》，21(10)：2833–2853。

Kieser, M. 和 Friede, T.（2000）。在控制 I 类错误率的情况下，重新计算内部预试验设计的样本量。Statistics in Medicine，19(7)：901–911。

Kieser, M. 和 Friede, T.（2003）。不影响 I 类错误率的盲态样本量调整简易程序。Statistics in Medicine，22(23)：3571–3581。

Kieser, M. 和 Wassmer, G.（1996）。关于在样本量确定中使用预试样本方差的置信上限。Biometrical Journal，38(8)：941–949。

Lehmann, E. L.（1955）。《有序分布族》。载于《E. L. Lehmann选集》，第757–777页。施普林格。

Miller, F.（2005）。临床研究中期样本量重新估计的方差估计。《Biometrics》，61(2)：355–361。

Placzek, M. and Friede, T. (2023). 适应性富集设计中的盲态样本量重新计算. Biometrical Journal, 65(2):2000345.

Schmidli, H., Neuenschwander, B., and Friede, T. (2017). 利用历史方差数据进行荟萃分析预测，以设计和分析临床试验. Computational Statistics & Data Analysis, 113:100–110.

Stein, C. (1945). 一种针对线性假设的双样本检验，其检验效能不依赖于方差. The Annals of Mathematical Statistics, 16(3):243–258.

Wittes, J., Schabenberger, O., Zucker, D., Brittain, E., and Proschan, M. (1999). 内部预试验 I：朴素 t 检验的 I 类错误率. Statistics in Medicine, 18(24):3481–3491.

Zucker, D. M., Wittes, J. T., Schabenberger, O., and Brittain, E. (1999). 内部预试验 II：各种程序的比较。Statistics in Medicine, 18(24):3493–3509.

表 1： 方差估计量汇总

Estimate	描述 iption	定义
$\hat{\sigma}^2_{OS}$	单样本方差	(2.1)
$\hat{\sigma}^2_{ADJ}$	调整后单样本方差	$\hat{\sigma}^2_{OS} - \delta^2 n_{1,\ int} n_{0,\ int}/(n_{int}(n_{int}-1))$
$\hat{\sigma}^2_{U,1-\gamma}$	不可行上置信限	$\hat{\sigma}^2_{OS}(n_{int}-1)/d_{\lambda,\ 1-\gamma}$
$\bar{\sigma}^2_{U,\ 1-\gamma}$	保守上置信限	$\hat{\sigma}^2_{OS}\ (n_{int}-1)\ /d_{1-\ \gamma}$

表2： 在均衡分配下进行盲态审查时，使用 (3.10) 推导出的置信水平 $1-\gamma$ $(n_{1,\ int} = n_{0,\ int})$ ，对应单侧显著性水平 $\alpha = 0.025$ 和目标检验效能 $1-\beta = 0.80$，0。90

$n_{z,int}$	2	3	4	5	6	7	8	9	10	20	30	40
$1-\beta = 0.80$	0.65	0.62	0.61	0.60	0.59	0.59	0.58	0.58	0.57	0.55	0.54	0.54
$1-\beta = 0.90$	0.76	0.72	0.69	0.67	0.66	0.65	0.64	0.63	0.62	0.59	0.57	0.57

Table 3: Power (%) with the re-calculated sample size based on the internal pilot study with $n_{z,int}$ subjects for the cases of $N = 32, 63$ and 174 ($\sigma^2 = 2.038, 4.013$ and 11.08); *One-sample*: (2.4) with (2.1), *IF*: (2.5), *Proposed*: (3.7), *Theoretical*: (3.3)

		Power			
N	$n_{z,int}$	*One-sample*	*IF*	*Proposed*	*Theoretical*
32	2	66.28	91.41	80.85	77.48
	3	71.15	85.92	81.01	77.28
	4	73.65	84.09	81.42	77.53
	5	75.17	83.28	81.53	77.53
	6	76.20	82.84	81.51	77.41
	7	76.96	82.57	81.42	77.25
	8	77.52	82.38	81.61	77.40
	9	77.97	82.26	81.46	77.20
	10	78.32	82.17	81.29	77.30
	20	80.02	82.88	81.66	77.32
	30	81.72	82.66	82.54	79.60
63	2	65.61	91.16	80.38	78.66
	3	70.45	85.46	80.45	78.53
	4	72.93	83.55	80.82	78.82
	5	74.44	82.68	80.90	78.83
	6	75.46	82.20	80.85	78.73
	7	76.20	81.90	80.74	78.58
	8	76.76	81.70	80.91	78.73
	9	77.20	81.56	80.74	78.53
	10	77.55	81.46	80.57	78.64
	20	79.17	81.08	80.86	78.56
	30	79.71	80.98	80.83	78.50
	40	79.99	80.94	80.92	78.58
	50	80.25	80.98	80.89	78.66
	60	81.19	81.65	81.62	80.06
174	2	65.31	91.04	80.15	79.53
	3	70.09	85.21	80.15	79.45
	4	72.53	83.23	80.47	79.74
	5	74.02	82.32	80.52	79.76
	6	75.02	81.81	80.45	79.67
	7	75.75	81.50	80.32	79.53
	8	76.29	81.28	80.48	79.68
	9	76.73	81.13	80.30	79.49
	10	77.08	81.02	80.12	79.60
	20	78.66	80.60	80.37	79.53
	30	79.12	80.46	80.33	79.47
	40	79.47	80.43	80.40	79.55
	50	79.63	80.40	80.30	79.44
	60	79.74	80.38	80.34	79.47
	75	79.85	80.36	80.36	79.50
	100	79.96	80.34	80.38	79.51
	125	80.03	80.33	80.30	79.43
	150	80.11	80.35	80.35	79.50

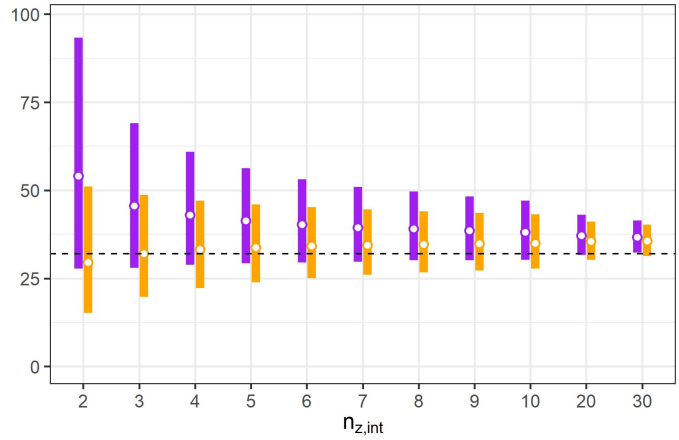
Table 4: Type I error rate (%) with the re-calculated sample size based on the internal pilot study with $n_{z,int}$ subjects for the cases of $N = 32, 63$ and 174 ($\sigma^2 = 2.038, 4.013$ and 11.08); *One-sample*: (2.4) with (2.1), *IF*: (2.5), *Proposed*: (3.7), *Theoretical*: (3.3)

		Type I error rate			
N	$n_{z,int}$	*One-sample*	*IF*	*Proposed*	*Theoretical*
32	2	2.416	2.453	2.433	2.429
	3	2.457	2.462	2.459	2.458
	4	2.471	2.473	2.472	2.471
	5	2.479	2.479	2.481	2.481
	6	2.485	2.485	2.487	2.484
	7	2.488	2.489	2.486	2.488
	8	2.494	2.492	2.491	2.493
	9	2.495	2.496	2.495	2.496
	10	2.495	2.492	2.495	2.494
	20	2.504	2.503	2.504	2.505
	30	2.505	2.505	2.505	2.505
63	2	2.435	2.470	2.453	2.449
	3	2.466	2.473	2.471	2.470
	4	2.476	2.480	2.478	2.477
	5	2.483	2.485	2.483	2.483
	6	2.488	2.489	2.488	2.488
	7	2.488	2.488	2.488	2.488
	8	2.492	2.495	2.492	2.493
	9	2.493	2.495	2.496	2.500
	10	2.492	2.491	2.491	2.492
	20	2.500	2.498	2.498	2.499
	30	2.501	2.500	2.501	2.501
	40	2.501	2.501	2.501	2.501
	50	2.503	2.502	2.503	2.503
	60	2.503	2.503	2.503	2.504
174	2	2.463	2.483	2.474	2.475
	3	2.481	2.487	2.485	2.485
	4	2.486	2.489	2.488	2.487
	5	2.489	2.491	2.491	2.490
	6	2.494	2.494	2.495	2.495
	7	2.493	2.494	2.492	2.493
	8	2.498	2.497	2.497	2.498
	9	2.498	2.498	2.498	2.499
	10	2.494	2.495	2.494	2.493
	20	2.497	2.497	2.497	2.498
	30	2.498	2.498	2.498	2.498
	40	2.500	2.499	2.499	2.499
	50	2.501	2.500	2.500	2.500
	60	2.500	2.500	2.500	2.500
	75	2.502	2.502	2.502	2.503
	100	2.500	2.501	2.500	2.501
	125	2.501	2.501	2.501	2.502
	150	2.502	2.502	2.503	2.502

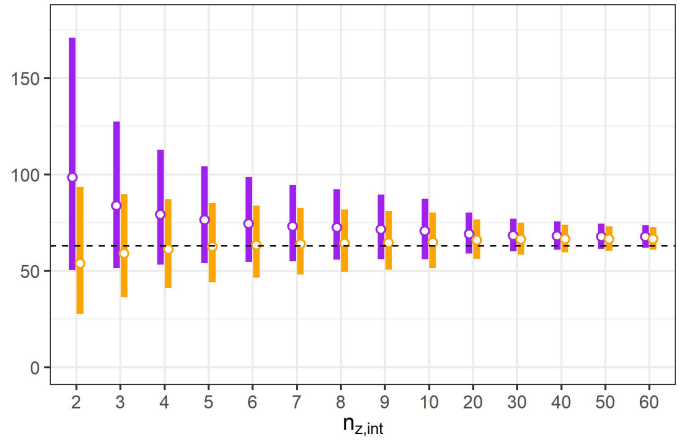
Table 5: Mean (standard deviation) of the re-calculated sample size based on the internal pilot study with $n_{z,int}$ subjects for the cases of $N = 32, 63$ and 174 ($\sigma^2 = 2.038, 4.013$ and 11.08); *One-sample*: (2.4) with (2.1), *IF*: (2.5), *Proposed*: (3.7), *Theoretical*: (3.3)

		Mean (standard deviation) of $\hat{n}_{z,fin}$			
N	$n_{z,int}$	*One-sample*	*IF*	*Proposed*	*Theoretical*
32	2	37.25 (30.09)	136.5 (110.3)	68.06 (55.01)	58.06 (46.92)
	3	36.72 (23.02)	64.64 (40.54)	52.11 (32.69)	45.23 (28.37)
	4	36.49 (19.35)	52.25 (27.72)	47.21 (25.04)	41.27 (21.89)
	5	36.36 (17.02)	47.29 (22.13)	44.48 (20.82)	39.05 (18.28)
	6	36.28 (15.36)	44.63 (18.90)	42.71 (18.08)	37.59 (15.92)
	7	36.23 (14.11)	42.98 (16.74)	41.43 (16.14)	36.54 (14.23)
	8	36.19 (13.12)	41.85 (15.18)	40.85 (14.82)	36.07 (13.08)
	9	36.16 (12.32)	51.03 (13.98)	40.03 (13.64)	35.39 (12.06)
	10	36.13 (11.64)	40.42 (13.03)	39.36 (12.69)	35.12 (11.32)
	20	36.04 (8.067)	37.96 (8.517)	37.73 (8.463)	33.53 (7.458)
	30	36.54 (5.815)	37.63 (6.212)	37.50 (6.167)	34.10 (4.699)
63	2	68.25 (55.55)	250.1 (203.6)	124.7 (101.5)	114.9 (93.54)
	3	67.71 (42.73)	119.2 (75.23)	96.12 (60.65)	89.32 (56.36)
	4	67.49 (35.99)	96.64 (51.54)	87.31 (46.57)	81.44 (43.43)
	5	67.36 (31.69)	87.61 (41.21)	82.41 (38.77)	77.02 (36.23)
	6	67.29 (28.63)	82.77 (35.22)	79.19 (33.70)	74.11 (31.54)
	7	67.23 (26.32)	79.76 (31.22)	76.88 (30.10)	72.01 (29.19)
	8	67.19 (24.49)	77.70 (28.32)	75.84 (27.64)	71.09 (25.91)
	9	67.16 (22.99)	76.22 (26.10)	74.36 (25.46)	69.73 (23.87)
	10	67.13 (21.74)	75.09 (24.32)	73.13 (23.68)	69.20 (22.41)
	20	67.03 (15.15)	70.61 (15.96)	70.18 (15.87)	65.96 (14.91)
	30	66.99 (12.31)	69.31 (12.74)	69.03 (12.69)	64.91 (11.93)
	40	66.98 (10.63)	68.69 (10.90)	68.64 (10.90)	64.57 (10.24)
	50	67.04 (9.335)	68.38 (9.569)	68.21 (9.540)	64.27 (8.796)
	60	67.90 (7.363)	68.84 (7.707)	68.77 (7.683)	65.60 (6.337)
174	2	179.3 (146.3)	657.0 (536.2)	327.6 (267.4)	317.9 (259.5)
	3	178.7 (113.0)	314.7 (199.0)	253.7 (160.4)	247.0 (156.2)
	4	178.5 (95.37)	255.6 (136.6)	230.9 (123.4)	225.1 (120.3)
	5	178.4 (84.06)	232.0 (109.3)	218.2 (102.8)	212.8 (100.3)
	6	178.3 (76.01)	219.3 (93.49)	209.8 (89.46)	204.8 (87.30)
	7	178.2 (69.89)	211.4 (82.92)	203.8 (79.93)	199.0 (78.03)
	8	178.2 (65.06)	206.1 (75.24)	201.1 (73.43)	196.4 (71.70)
	9	179.2 (61.10)	202.2 (69.34)	197.3 (67.65)	192.7 (66.07)
	10	179.1 (57.79)	199.3 (64.63)	194.1 (62.95)	191.2 (62.01)
	20	178.0 (40.31)	187.6 (42.47)	186.4 (42.21)	182.2 (41.25)
	30	178.0 (32.76)	184.1 (33.90)	183.4 (33.76)	179.3 (33.00)
	40	178.0 (28.31)	182.5 (29.04)	182.4 (29.02)	178.3 (28.37)
	50	178.0 (25.29)	181.6 (25.80)	181.1 (25.74)	177.1 (25.16)
	60	178.0 (23.07)	180.9 (23.45)	180.7 (23.43)	176.7 (22.90)
	75	178.0 (20.61)	180.3 (20.89)	180.3 (20.89)	176.3 (20.42)
	100	177.9 (17.84)	179.7 (18.01)	179.9 (18.03)	175.9 (17.63)
	125	177.9 (15.94)	179.3 (16.07)	179.2 (16.06)	175.3 (15.70)
	150	178.1 (14.31)	179.2 (14.45)	179.2 (14.44)	175.3 (13.96)

(a) $N = 32$



(b) $N = 63$



(c) $N = 174$

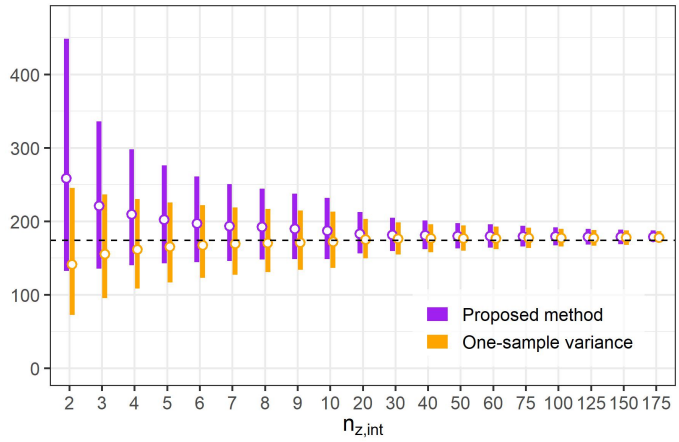


图1：在均衡分配下，采用所提出的方法和单样本方法重新计算的$\hat{n}_{z,\ fin}$的分布（中位数及1^{st}和3^{rd}四分位数），适用于以下情形：(a) $N = 32$、(b) $N = 63$和(c) $N = 174$（$n_{1,\ int} = n_{0,\ int}$，$\pi = 0.5$）。