标准差检验（两个或更多样本）

概述

Minitab 协助包含两种用于比较独立样本以确定其变异性是否存在显著差异的分析。双样本标准差检验可比较两个样本的标准差，标准差检验可比较两个以上样本的标准差。在本白皮书中，我们将 k = 2 的 k 样本设计称作双样本设计，将 k > 2 的 k 样本设计称作多样本设计。通常，我们将分别对这两种类型的设计进行研究（请参见附录 A）。

由于标准差是方差的平方根，因此，用于比较标准差的假设检验等效于用于比较方差的假设检验。我们开发了许多统计方法，来比较来自两个或更多总体的方差。在这些检验中，Levene/Brown-Forsythe 检验是稳健度最高且最常用的检验之一。但是，Levene/Brown-Forsythe 检验的功效性能的满意度不如其在双样本设计中的类型 I 误差属性。Pan (1999) 指出，对于某些总体（包括正态总体），双样本设计中的检验功效具有一个远低于 1 的上限（与标准差之间差值的量值无关）。换言之，对于这些类型的数据，此检验得出标准差之间不存在差值的结论的可能性更高（与差值大小无关）。鉴于上述原因，“协助”将为双样本标准差检验使用新的检验，即 Bonett 检验。对于采用多样本设计的标准差检验，“协助”将使用多重比较 (MC) 过程。

Bonett (2006) 检验（即 Layard (1978) 的双方差等同性检验的修订版本）可提高小样本的检验性能。Banga 和 Fox (2013A) 可推导出与 Bonett 检验关联的置信区间，显示其准确度等同于与 Levene/Brown-Forsythe 检验关联的置信区间，并且比大多数分布更准确。此外，Banga 和 Fox (2013A) 确定了 Bonett 检验与 Levene/Brown-Forsythe 检验的稳健度相同，并且比大多数分布的功能更强大。

多重比较 (MC) 过程包含多个样本的标准差（或方差）的同质性或等同性的总体检验，它基于每对标准差的比较区间。推导出比较区间，以便当且仅当至少有一对比较区间不重叠时，MC 检验才显著。Banga 和 Fox (2013B) 指出，MC 检验具有类型 I 和类型 II 误差属性，这些属性与大多数分布的 Levene/Brown-Forsythe 检验类似。MC 检验的一个重要优势是比较区间的图形显示，它提供用于找出具有不同标准差的样本的高效可视工具。当设计中只有两个样本时，MC 检验等效于 Bonett 检验。

在本白皮书中，我们针对不同的数据分布和样本数量评估了 Bonett 检验和 MC 检验的有效性。此外，我们还调查了用于 Bonett 检验的功效和样本数量分析，该调查基于大样本近似方法进

行。根据这些因素，我们开发了“协助”可自动对您的数据执行并在“报告卡”中显示的以下检查方法：

- 异常数据
- 正态性
- 检验的有效性
- 样本数量（仅限双样本标准差检验）

标准差检验方法

Bonett 检验和 MC 检验的有效性

在等方差（Conover 等人， 1981）的检验比较研究中发现，Levene/Brown-Forsythe 检验是具有最佳执行效果的检验之一（根据其类型 I 和类型 II 误差率）。此后，还建议采用其他方法检验双样本和多样本设计中的等方差（Pan，1999；Shoemaker，2003；Bonett，2006）。例如，Pan 指出，尽管 Levene/Brown-Forsythe 检验具有稳健性和解释简单的特点，但在样本源自某些总体（包括正态总体）时，它没有足够的功效来检测两个标准差之间的重要差值。由于这一关键限制，“协助”将对双样本标准差检验使用 Bonett 检验（请参见附录 A 或 Banga 和 Fox，2013A）。对于具有两个以上样本的标准差检验，“协助”将结合使用 MC 过程和比较区间，从而可在 MC 检验效果显著时，提供一个图形显示，以确定具有不同标准差的样本（请参见附录 A 或 Banga 和 Fox，2013B）。

目标

首先，我们想要在比较两个总体标准差时评估 Bonett 检验的性能。其次，我们想要在比较两个以上总体的标准差时评估 MC 检验的性能。尤其是，我们想要在对来自不同类型的分布的不同数量的样本执行检验时，评估这些检验的有效性。

方法

附录 A 中定义了用于 Bonett 检验和 MC 检验的统计方法。为评估这些检验的有效性，我们需要检查其类型 I 误差率在不同条件下是否仍继续接近目标显著性水平（alpha 值）。为此，我们进行了一组模拟，以在比较两个独立样本的标准差时评估 Bonett 检验的有效性，并进行了其他多组模拟，以在比较多个 (k) 独立样本（k > 2）的标准差时评估 MC 检验的有效性。

通过使用平衡与不平衡设计，我们根据几种分布生成了 10,000 对或多个 (k) 不同数量的随机样本。然后，我们使用目标显著性水平 $\alpha = 0.05$，执行了双侧 Bonett 检验，以比较两个样本的标准差，或执行了 MC 检验，以比较每个试验中 k 个样本的标准差。我们从 10,000 个仿行中计算了检验否定原假设的次数（实际上此时真实的标准差相等），并将此比例（即模拟显著性水平）与目标显著性水平进行比较。如果检验的执行效果不错，则表示实际类型 I 误差率的模拟显著性水平应非常接近目标显著性水平。有关用于双样本和 k 样本模拟的特定方法的详细信息，请参见附录 B。

结果

对于双样本比较，当样本数量中等或较大时，Bonett 检验的模拟类型 I 误差率接近目标显著性水平（与分布无关，也与设计是否平衡设计无关）。但是，在从极度偏斜的总体抽取小样本时，Bonett 检验通常比较保守，它具有略低于目标显著性水平（即，目标类型 I 误差率）的类型 I 误差率。

对于多样本比较，当样本数量中等或较大时，MC 检验的类型 I 误差率接近目标显著性水平（与分布无关，也与设计是否平衡设计无关）。但是，对于小样本和极度偏斜样本，Bonett 检验通常不太保守，当设计中的样本数较大时，类型 I 误差率略高于目标显著性水平。

我们的研究结果与 Banga 和 Fox (2013A) 和 (2013B) 的研究结果保持一致。我们得出以下结论：在最小样本数量至少为 20 时，Bonett 检验和 MC 检验的执行效果不错。因此，我们

使用协助报告卡中检验“有效性”检查中的这一最小样本数量要求（请参见“数据检查”部分）。

比较区间

如果用于比较两个或多个标准差的检验统计意义显著，表明其中至少有一个标准差与其他标准差不同，则分析的下一步是确定哪些样本在统计意义上显著不同。进行此比较的直观方法是绘制与每个样本关联的置信区间图形，然后找出其区间不重叠的样本。但是，根据图形得出的结论可能与检验结果不符，原因是没有为比较设计对应的置信区间。

目标

我们想要开发一种计算各个比较区间的方法，以用作方差同质性的总体检验方法并用于在总体检验显著时，找出具有不同方差的样本。MC 过程的关键要求如下：当且仅当至少有一对比较区间不重叠（表示至少有两个样本的标准差不同）时，总体检验效果才显著。

方法

我们用于比较多个标准差的 MC 过程根据多重配对比较推导而出。通过使用两个总体标准差的等同性的 Bonett (2006) 检验比较每对样本。根据 Nayakama (2009) 中所示的大样本近似方法，配对比较使用多重性校正。由于常用 Bonferroni 校正会随着样本数的增加而变得越来越保守，因此，Bonferroni 校正时最好使用大样本近似方法。最后，根据 Hochberg 等人 (1982) 的最佳近似过程，从配对比较中得出比较区间。有关详细信息，请参见附录 A。

结果

MC 过程可满足以下要求：当且仅当至少有两个比较区间不重叠时，标准差的等同性的总体检验效果才显著。如果整体检验不显著，则所有比较区间必须重叠。

“协助”会在汇总报告的“标准差比较图”中显示比较区间。在此图旁边，“协助”将显示 MC 检验的 p 值，该检验是标准差同质性的总体检验。在标准差检验统计意义显著时，与至少一个其他区间不重叠的任何比较区间会用红色标记出来。如果标准差检验统计意义不显著，则任何区间都不会用红色标记出来。

理论功效的性能（仅限双样本设计）

要计划样本数量，需要使用 Bonett 和 MC 检验的理论功效函数。对于双样本设计，可使用大样本理论方法推导出检验的近似理论功效函数。由于此函数源自大样本近似方法，因此，我们需要在使用根据正态和非正态分布生成的小样本进行检验时评估其属性。但是，当我们比较两组以上的标准差时，获得 MC 检验的理论功效函数并不容易。

目标

我们想要确定能否根据大样本近似方法，使用理论功效函数在“协助”中评估双样本标准差检验的功效和样本数量要求。为此，我们需要评估近似理论功效函数是否精确反映对来自多种类型的分布（包括正态和非正态分布）的数据执行 Bonett 检验时，该检验所取得的实际功效。

方法

附录 C 中推导出了适用于双样本设计的 Bonett 检验的近似理论功效函数。

我们使用 Bonett 检验执行了用于估计实际功效水平的模拟（我们将其称作模拟功效水平）。首先，我们根据几种分布（包括正态分布和非正态分布）生成了不同数量的随机样本对。对于每种分布，我们对 10,000 对样本仿行中的每一对仿行执行了 Bonett 检验。对于每对样本数量，我们计算了检验的模拟功效，以检测作为检验效果显著的 10,000 对样本的一部分的给定差值。要进行比较，我们还使用此检验的近似理论功效函数计算了对应的功效水平。如果近似方法效果不错，则理论和模拟功效水平应该比较接近。有关详细信息，请参见附录 D。

结果

我们的模拟表明，对于大多数分布，针对小数量样本的 Bonett 检验的理论和模拟功效函数几乎相等，并在最小样本数量达到 20 时更为接近。对于轻尾到中尾的对称和近对称分布，理论功效水平略高于模拟（实际）功效水平。但是，对于偏斜分布和重尾分布，理论功效水平略低于模拟（实际）功效水平。有关详细信息，请参见附录 D。

总之，我们的结果表明，理论功效函数为计划样本数量提供了良好的基础。

数据检查

异常数据

异常数据是极大或极小的数据值，也称作异常值。异常数据会对分析结果产生较大的影响，并且会影响找到统计意义显著的结果的几率，尤其是在样本比较小时。异常数据表示数据收集出现问题，或者可能由您所研究的过程中的异常行为导致。因此，这些数据点通常值得调查，并在可能时给予校正。模拟研究显示，当数据包含异常值时，Bonett 检验和 MC 检验比较保守（参见附录 B）。检验的实际显著性水平明显低于目标显著性水平，尤其是在使用小样本进行分析时。

目标

我们想要开发一种方法，用于检查相对于总体样本而言非常大或非常小，并且会影响分析结果的数据值。

方法

我们根据 Hoaglin、Iglewicz 和 Tukey (1986) 描述的用于在箱线图中找出异常值的方法开发了用于检查异常值的方法。

结果

如果某个数据点超过分布的下四分位或上四分位的四分位间距的 1.5 倍，则“协助”就会将该数据点标识为异常值。下四分位和上四分位是数据的第 25 个和第 75 个四分位。四分位间距是两个四分位之间的差值。即使存在多个异常值，此方法的效果也不错，因为它可以检测到每个特定的异常值。

在检查异常数据时，“协助”会在“报告卡”中显示以下状态指示符：

状态	条件
	不存在异常数据点。
	至少有一个数据点异常，可能会对结果造成较大的影响。

正态性

与在正态性假设条件下推导出的大多数方差等同性检验不同的是，针对标准差等同性的 Bonett 检验和 MC 检验不会假设特定的数据分布。

目标

虽然 Bonett 检验和 MC 检验基于大样本近似方法，但我们想要确认对小样本的正态和非正态数据执行这些检验的效果。我们还想告诉用户数据正态性与标准差检验结果的相关情况。

方法

要在不同条件下评估检验的有效性，我们使用不同样本数量的正态和非正态数据进行了模拟，以检查 Bonett 检验和 MC 检验的类型 I 误差率。有关详细信息，请参见标准差检验方法部分和附录 B。

结果

我们的模拟表明，只要样本数量足够大（最小样本数量 ≥ 20），数据分布不会对 Bonett 检验或 MC 检验的类型 I 误差属性产生重要影响。这些检验产生的类型 I 误差率始终接近正态和非正态数据的目标误差率。

根据这些有关类型 I 误差率的结果，“协助”会在“报告卡”中显示有关正态性的信息。

对于双样本设计，“协助”会显示以下指示符：

状态	条件
	此分析使用 Bonett 检验。如果样本足够大，对正态数据和非正态数据执行检验的效果都不错。

对于多样本设计，“协助”显示以下指示符：

状态	条件
	此分析使用多重比较检验。如果样本足够大，对正态数据和非正态数据执行检验的效果都不错。

检验的有效性

在标准差检验方法部分，我们解释了对于双样本和多重 (k) 比较，当样本数量为中等或较大时，Bonett 检验和 MC 检验产生的类型 I 误差率接近平衡和不平衡设计中正态和非正态数据的目标误差率。但是，当样本为小样本时，Bonett 和 MC 检验通常效果不佳。

目标

我们想要应用一项规则，以根据用户数据，针对双样本和多 (k) 样本评估标准差检验结果的有效性。

方法

为在不同条件下评估检验的有效性，我们根据各种数据分布、样本数和样本数量执行了模拟，以检查 Bonett 检验和 MC 检验的类型 I 误差率，如前面的标准差检验方法部分中所述。有关详细信息，请参见附录 B。

结果

当最小样本数量至少为 20 时，Bonnet 检验和 MC 检验的执行效果不错。因此，“协助”会在“报告卡”中显示以下状态指示符，以评估标准差检验的有效性。

状态	条件
	样本数量至少为 20，因此，p 值应是准确的。
	有些样本数量小于 20，因此，p 值可能不准确。请考虑将样本数量至少增加到 20。

样本数量（仅适用于双样本标准差检验）

通常，为收集否定“无差异”原假设的证据，将执行统计假设检验。如果样本太小，检验的功效可能不准确，因此检测不到是否真正存在差值，这将导致类型 II 误差。因此，一定要确保样本数量足够大，以便有较高的概率检测到实际的重要差值。

目标

如果数据没有提供足够的证据来否定原假设，则我们想要确定样本数量是否足够大，以便有较高的概率检测到实际的重要差值。虽然计划样本数量的目的是确保样本数量足够大，以便有较高的概率检测到重要差值，但它们不应该大到有较高的概率使无意义的差值变成具有显著的统计意义。

方法

双样本标准差检验的功效和样本数量分析基于 Bonett 检验的近似功效函数，这通常提供该检验的实际功效函数的准确估计值（请参见方法部分的理论功效函数性能中汇总的模拟结果）。

结果

在数据不能提供足够的证据否定原假设时，“协助”将使用 Bonett 检验的近似功效函数来计算通过 80% 和 90% 的概率检测到的给定样本数量的实际差值。此外，如果用户提供了相关的特定实际差值，则“协助”将使用正态近似检验的功效函数来计算有 80% 和 90% 的机会检测到差值的样本数量。

为帮助解释结果，“协助”的双样本标准差检验的“报告卡”会在检查功效和样本数量时显示以下状态指示符：

状态	条件
	此检验发现标准差之间存在差值，因此，功效不是问题。 或 功效足够。此检验没有发现标准差之间存在差值，但样本数量足够大，至少有 90% 的机会检测到给定差值。
	功效可能足够。此检验没有发现标准差之间存在差值，但样本数量足够大，有 80% 到 90% 的机会检测到给定差值。将报告获取 90% 的功效所需的样本数量。

状态	条件
	功效可能不足。此检验没有发现标准差之间存在差值，但样本数量足够大，有 60% 到 80% 的机会检测到给定差值。将报告获取 80% 和 90% 的功效所需的样本数量。
	功效不足。此检验没有发现标准差之间存在差值，并且样本数量不够大，无法提供 60% 的机会检测到给定差值。将报告获取 80% 和 90% 的功效所需的样本数量。
	此检验没有发现标准差之间存在差值。您没有指定要检测的实际差值，因此，此报告根据您的样本数量和 alpha 值指出有 80% 和 90% 的机会检测到的差值。

参考书

Arnold, S.F. (1990). *Mathematical statistics*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Banga, S.J. and Fox, G.D. (2013A). 关于 Bonett 的标准差比率的稳健性置信区。*白皮书，Minitab Inc*.

Banga, S.J. and Fox, G.D. (2013B) 多个标准差的图形化多重比较过程。*白皮书，Minitab Inc*.

Bonett, D.G. (2006). 标准差比率的稳健性置信区间。*Applied Psychological Measurements*, 30, 432-439.

Brown, M.B., & Forsythe, A.B. (1974)。方差等同性的稳健性检验。*Journal of the American Statistical Association*, *69*, 364-367.

Conover, W.J., Johnson, M.E., & Johnson, M.M. (1981)。方差同质性检验的比较研究以及外大陆架招标数据的应用。*Technometrics, 23*, 351-361.

Gastwirth, J. L. (1982)。纳税评估均匀性度量的统计属性。*Journal of Statistical Planning and Inference, 6*, 1-12.

Hochberg, Y., Weiss G., and Hart, S. (1982)。关于多重比较的图形化过程。*Journal of the American Statistical Association, 77,* 767-772.

Layard, M.W.J. (1973)。方差同质性的稳健性大样本检验。*Journal of the American Statistical Association*, 68, 195-198.

Levene, H. (1960)。方差等同性的稳健性检验。In I. Olkin (Ed.), *Probability and statistics* (278-292). Stanford University Press, Palo Alto, California.

Nakayama, M.K. (2009)。异步有效的单阶段多重比较过程。*Journal of Statistical Planning and Inference*, 139, 1348-1356.

Pan, G. (1999)。关于双方差等同性的 Levene 类型检验。*Journal of Statistical Computation and Simulation*, 63, 59-71.

Shoemaker, L. H. (2003)。解决等方差的 F 检验问题。*The American Statistician, 57 (2)*, 105-114。

附录 A：Bonett 检验和多重比较检验的方法

利用 Bonett 方法（双样本设计）或多重比较（MC）过程（多样本设计）进行有关标准差或方差的推断的基本假设描述如下。设 $X_{11},\ldots,X_{1n_1},\ldots,X_{k1},\ldots,X_{kn_k}$ 为 k（$k \geq 2$）独立随机样本，$i = 1,\ldots,k$ 的每个样本分别提取自具有未知均值 μ_i 和方差 σ_i^2 的分布。假设样本的父分布具有常用有限峰度 $\gamma = E(Y-\mu)^4/\sigma^4 < \infty$。虽然此假设对理论推导至关重要，但对于样本数量足够大的大多数实际应用而言却不重要（Banga 和 Fox，2013A）。

方法 A1：双方差等同性的 Bonett 检验

Bonett 检验仅适用于比较两个方差或标准差的双样本设计。此检验是双样本设计中方差等同性的 Layard（1978）检验的修订版本。当且仅当满足以下条件时，显著性水平为 α 的双方差等同性的双侧 Bonett 检验才会否定等同性原假设，

$$\left|\ln(c\, S_1^2/S_2^2)\right| > z_{\alpha/2}\sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

其中：

S_i 是样本 i 的样本标准差

$g_i = (n_i - 3)/n_i, i = 1,2$

$z_{\alpha/2}$ 指标准正态分布的上 $\alpha/2$ 四分位

$\hat{\gamma}_P$ 是给定为以下值的合并峰度估计量：

$$\hat{\gamma}_P = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(X_{1j} - m_1)^4 + \sum_{j=1}^{n_2}(X_{2j} - m_2)^4}{[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]^2}$$

在合并峰度估计量的表达式中，m_i 是样本 i 的调整均值，调整比例为 $1/[2(n_i - 4)^{1/2}]$。

按上述，包括作为小样本调整的常数 c，以降低不平衡设计中不等尾误差概率的影响。此常数给定为 $c = c_1/c_2$，其中，

$$c_{\mathrm{i}} = \frac{n_{\mathrm{i}}}{n_{\mathrm{i}} - z_{\alpha/2}}, i = 1,2$$

如果设计是平衡设计，即如果使用 $n_1 = n_2$，则获得的检验的 p 值如下所示

$$\mathrm{P} = 2\,\mathrm{Pr}(Z > z)$$

其中，Z 是一个随机变量，该变量根据现有数据作为以下统计量的标准正态分布和 z 观测值分布。该统计量为

$$Z = \frac{\ln(C\, S_1^2/S_2^2)}{se}$$

其中

$$se = \sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

此外，如果设计为不平衡设计，则获得的检验的 p 值如下所示

$$P = 2\min(\alpha_L, \alpha_U)$$

其中，$\alpha_L = \Pr(Z > z_L)$ 和 $\alpha_U = \Pr(Z > z_U)$。变量 z_L 是函数的最小根

$$L(z, S_1, S_2, n_1, n_2) = \ln\frac{n_1}{n_2} + \ln\frac{n_2 - z}{n_1 - z} - z\,se + \ln\frac{S_1^2}{S_2^2} - \ln\rho_o^2, z < \min(n_1, n_2)$$

并且，z_U 是函数 $L(z, S_2, S_1, n_2, n_1)$ 的最小根。

方法 A2：多重比较检验和比较区间

假设存在 k $(k \geq 2)$ 个独立组或样本。我们的目标是为总体标准差构造一个 k 区间系统，以便当且仅当至少有两个 k 区间不重叠时，标准差等同性的检验效果才显著。这些区间也称作比较区间。此比较方法类似于单因子方差分析模型中均值的多重比较过程，这种模型最初是由 Tukey-Kramer 开发的，后来由 Hochberg 等人 (1982) 进行了普及。

比较两个标准差

对于双样本设计，可以直接计算与 Bonett 检验关联的标准差比率的置信区间，以评估标准差之间的差值大小（Banga 和 Fox，2013A）。实际上，我们对 Minitab 版本 17 中的“统计”>“基本统计量”>“双方差”使用此方法。但在“协助”中，我们想要提供比标准差比率的置信区间更容易解释的比较区间。为此，我们使用了方法 A1 中描述的 Bonett 过程来确定双样本的比较区间。

如果存在两个样本，当且仅当与方差等同性的 Bonett 检验关联的以下验收区间不包含 0 时，方差等同性的 Bonett 检验才效果显著。

$$\ln(c_1 S_1^2) - \ln(c_2 S_2^2) \pm z_{\alpha/2}\sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

其中，合并峰度估计值 $\hat{\gamma}_P$ 和 $g_i, i = 1,2$ 采用以前给定的值。

从此区间中，我们可以断定以下两个比较区间，以便当且仅当这两个区间不重叠时，方差或标准差等同性的检验效果才显著。这两个区间是

$$\left[S_i\sqrt{C_i\exp(-z_{\alpha/2}V_i)}, S_i\sqrt{C_i\exp(z_{\alpha/2}V_i)}\right], i = 1,2$$

其中

$$V_i = \frac{\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1}}}{\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1}} + \sqrt{\frac{\hat{\gamma}_P - g_j}{n_j - 1}}}\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1} + \frac{\hat{\gamma}_P - g_j}{n_j - 1}}, i = 1,2; j = 1,2; i \neq j$$

将这些区间用作标准差等同性的检验过程等效于标准差等同性的 Bonett 检验。尤其是，当且仅当标准差等同性的 Bonett 检验效果显著时，这些区间才不会重叠。但请注意，这些区间并

不是标准差置信区间，仅适用于标准差的多重比较。出于相同的原因，Hochberg 等人将比较均值的类似区间称作不确定性区间。我们将这些区间称作比较区间。

由于比较区间过程等效于标准差等同性的 Bonett 检验，因此与比较区间关联的 p 值与前面所述的双标准差等同性的 Bonett 检验的 p 值相同。

比较多个标准差

如果存在两个以上的组或样本，则可根据标准差等同性的 $k(k-1)/2$ 配对同时检验以及全族显著性水平 α 断定 k 比较区间。更具体地说，设 $X_{i1},\ldots,X_{in_i}$ 和 $X_{j1},\ldots,X_{jn_j}$ 为任何样本对 (i,j) 的样本数据。特定样本对 (i,j) 的标准差等同性检验类似于双样本情况，当且仅当区间不包含 0 时，达到某个 α' 水平的检验效果才显著。

$$\ln(c_i S_i^2) - \ln(c_j S_j^2) \pm z_{\alpha'/2}\sqrt{\frac{\hat{\gamma}_{ij}-g_i}{n_i-1}+\frac{\hat{\gamma}_{ij}-g_j}{n_j-1}}$$

如上所述，$\hat{\gamma}_{ij}$ 是基于样本对 (i,j) 的合并峰度估计量，并且给定为

$$\hat{\gamma}_{ij} = (n_i+n_j)\frac{\sum_{l=1}^{n_i}(X_{il}-m_i)^4+\sum_{l=1}^{n_j}(X_{jl}-m_j)^4}{\left[(n_i-1)S_i^2+(n_j-1)S_j^2\right]^2}$$

此外，如上所定义，m_i 是样本 i 的调整均值，调整比例为 $1/[2(n_i-4)^{1/2}]$，并且

$$g_i=\frac{n_i-3}{n_i}, g_j=\frac{n_j-3}{n_j}, c_i=\frac{n_i}{n_i-z_{\alpha'/2}}, c_j=\frac{n_j}{n_j-z_{\alpha'/2}}$$

由于存在 $k(k-1)/2$ 个同时配对检验，因此，必须选择水平 α'，实际全族误差率才接近目标显著性水平 α。一次可能的调整基于 Bonferroni 的近似方法。但是，已知 Bonferroni 校正会随着设计中样本数的增加而越来越保守。更好的方法基于 Nakayama (2008) 给定的正态近似方法。通过使用此方法，我们只能使用 $q_{\alpha,k}/\sqrt{2}$ 替换 $z_{\alpha'/2}$，其中，$q_{\alpha,k}$ 是 k 独立和相同分布标准正态随机变量系列的上限点 α，即，

$$\Pr\left(\max_{1\le i<j\le k}|Z_i-Z_j|\le q_{\alpha,k}\right)=1-\alpha$$

其中，$Z_1,\ldots,Z_k$ 是独立和相同分布标准正态随机变量。

此外，使用类似于 Hochberg 等人 (1982) 提供的方法，当且仅当某个样本对 (i,j) 满足以下条件时，最接近上述配对过程的过程将否定标准差等同性的原假设

$$\left|\ln(c_i S_i^2)-\ln(c_j S_j^2)\right| > q_{\alpha,k}(V_i+V_j)/\sqrt{2}$$

其中，选择 V_i 可使等同性最小

$$\sum_{i\ne j}\sum\left(V_i+V_j-b_{ij}\right)^2$$

并且

$$b_{ij}=\sqrt{\frac{\hat{\gamma}_{ij}-g_i}{n_i-1}+\frac{\hat{\gamma}_{ij}-g_j}{n_j-1}}$$

Hochberg 等人 (1982) 中所述的此问题的解决方法是选择

$$V_i = \frac{(k-1)\sum_{j\neq i} b_{ij} - \sum\sum_{1\leq j<l\leq k} b_{jl}}{(k-1)(k-2)}$$

它遵循以下原则：当且仅当以下 k 区间中至少有一对不重叠时，基于近似过程的检验才效果显著。

$$\left[S_i\sqrt{\mathrm{C_i}\exp(-q_{\alpha,k}V_i/\sqrt{2}\,)}, S_i\sqrt{C_i\exp(q_{\alpha,k}V_i/\sqrt{2})}\,\right], i = 1, \ldots, k$$

为计算与 MC 检验关联的总体 p 值，我们设 P_{ij} 为与任何样本对 (i,j) 关联的 p 值。然后，它遵循以下原则：与多重比较检验关联的总体 p 值是

$$P = \min\{P_{ij}, 1 \leq i < j \leq k\}$$

为计算 P_{ij}，我们使用以下方式，执行了方法 A1 中给定的双样本设计的算法：

$$se = V_i + V_j$$

其中，V_i 为如上给定的值。

具体地说，如果 $n_i \neq n_j$

$$P_{ij} = \min(\alpha_L, \alpha_U)$$

其中 $\alpha_L = \Pr(Q > z_L\sqrt{2})$、$\alpha_U = \Pr(Q > z_U\sqrt{2})$，变量 z_L 是函数 $L(z, S_i, S_j, n_i, n_j)$ 的最小根，变量 z_U 是函数 $L(z, S_j, S_i, n_j, n_i)$ 的最小根，并且 Q 是具有前面所定义的范围分布的随机变量。

如果 $n_i = n_j$，则 $P_{ij} = \Pr(Q > |z_o|\sqrt{2})$，其中，

$$z_o = \frac{\ln S_i^2 - \ln S_j^2}{V_i + V_j}$$

附录 B：Bonett 检验和多重比较检验的有效性

模拟 B1：Bonett 检验的有效性（双样本模型，平衡和不平衡设计）

我们根据具有不同属性的分布生成了一对样本数量为小到中等的随机样本。分布包括：

- 标准正态分布 (N(0,1))
- 对称和轻尾分布，包括均匀分布 (U(0,1)) 和 Beta 分布，这两个参数均设置为 3 (B(3,3))
- 对称和重尾分布，包括自由度分别为 5 和 10 的 t 分布 (t(5) 和 t(10)) 以及位置为 0 和尺度为 1 的 Laplace 分布 (Lpl)
- 偏斜和重尾分布，包括尺度为 1 的指数分布 (Exp) 和自由度分别为 5 和 10 的卡方分布 (Chi(5) 和 Chi(10))
- 左偏斜和重尾分布；尤其是参数分别设置为 8 和 1 的 Beta 分布 (B(8,1))

此外，为了评估异常值的直接影响，我们根据定义如下的污染正态分布生成了样本对：

$$CN(p,\sigma) = pN(0,1) + (1-p)N(0,\sigma)$$

其中，p 是混合参数，$1-p$ 是污染比例（等于异常值比例）。我们为研究选择了两个污染正态总体：$CN(0.9,3)$，其中，10% 的总体为异常值；$CN(0.8,3)$，其中，20% 的总体为异常值。这两种分布对称，并因存在异常值而呈长尾分布。

我们对每种分布中的每对样本执行了目标显著性水平为 $\alpha = 0.05$ 的双侧 Bonett 检验。由于在每种情况下，模拟的显著性水平都基于 10,000 对样本仿行，并且我们使用了目标显著性水平 5%，因此，模拟误差为 $\sqrt{0.95(0.05)/10{,}000} = 0.2\%$。

模拟结果汇总如下表 1 中所示。

表 1　在平衡和不平衡双样本设计中，双侧 Bonett 检验的模拟显著性水平。目标显著性水平为 0.05。

分布	n_1, n_2	模拟水平	分布	n_1, n_2	模拟水平
N(0,1)	10，10	0.038	Exp	10，10	0.052
	20，10	0.043		20，10	0.051
	20，20	0.045		20，20	0.049
	30，10	0.044		30，10	0.044
	30，20	0.046		30，20	0.042
	25，25	0.048		25，25	0.043

分布	n_1, n_2	模拟水平	分布	n_1, n_2	模拟水平
	30，30	0.048		30，30	0.042
	40，40	0.051		40，40	0.042
	50，50	0.047		50，50	0.039
t(5)	10，10	0.044	Chi(5)	10，10	0.040
	20，10	0.042		20，10	0.043
	20，20	0.046		20，20	0.040
	30，10	0.041		30，10	0.039
	30，20	0.046		30，20	0.043
	25，25	0.048		25，25	0.042
	30，30	0.043		30，30	0.043
	40，40	0.046		40，40	0.040
	50，50	0.050		50，50	0.039
t(10)	10，10	0.041	Chi(10)	10，10	0.044
	20，10	0.040		20，10	0.042
	20，20	0.045		20，20	0.041
	30，10	0.046		30，10	0.043
	30，20	0.045		30，20	0.045
	25，25	0.046		25，25	0.046
	30，30	0.048		30，30	0.038
	40，40	0.045		40，40	0.042
	50，50	0.051		50，50	0.049
Lpl	10，10	0.054	B(8,1)	10，10	0.053
	20，10	0.056		20，10	0.045
	20，20	0.055		20，20	0.048
	30，10	0.057		30，10	0.042
	30，20	0.058		30，20	0.047
	25，25	0.057		25，25	0.041

分布	n_1, n_2	模拟水平	分布	n_1, n_2	模拟水平
	30，30	0.053		30，30	0.040
	40，40	0.047		40，40	0.042
	50，50	0.048		50，50	0.038
B(3,3)	10，10	0.032	CN(0.9,3)	10，10	0.024
	20，10	0.037		20，10	0.022
	20，20	0.042		20，20	0.018
	30，10	0.039		30，10	0.019
	30，20	0.038		30，20	0.020
	25，25	0.039		25，25	0.019
	30，30	0.041		30，30	0.015
	40，40	0.044		40，40	0.020
	50，50	0.046		50，50	0.017
U(0,1)	10，10	0.030	CN(0.8,3)	10，10	0.022
	20，10	0.032		20，10	0.019
	20，20	0.031		20，20	0.020
	30，10	0.034		30，10	0.017
	30，20	0.034		30，20	0.020
	25，25	0.034		25，25	0.021
	30，30	0.037		30，30	0.017
	40，40	0.043		40，40	0.023
	50，50	0.043		50，50	0.020

如表 1 所示，对于轻尾到中尾的对称或近对称分布，当样本数量较小时，Bonett 检验的模拟显著性水平低于目标显著性水平 (0.05)。另一方面，当小样本源自高度偏斜的分布时，模拟水平倾向于略大于目标水平。

当样本数量中等或较大时，模拟显著性水平接近所有分布的目标显著性水平。实际上，检验效果相当好，即使是高度偏斜的分布（如指数分布和 Beta(8,1) 分布）也不例外。

此外，异常值对小样本的影响似乎比大样本大。当双样本的最小数量达到 20 时，污染正态总体的模拟显著性水平将大致稳定在 0.020。

当双样本的最小数量达到 20 时，模拟显著性水平全部落入区间 [0.038, 0.058]，但扁平均匀分布和污染正态分布除外。虽然对于目标显著性水平 0.05 而言，模拟显著性水平 0.040 略为保守，但对于大多数实际用途，此类型 I 误差率是可以接受的。因此，我们得出结论，当双样本最小数量至少为 20 时，Bonett 检验有效。

模拟 B2：MC 检验的有效性（多样本模型）

部分 I：平衡设计

我们进行了模拟，以在采用平衡设计的多样本模型中检查 MC 检验的性能。我们使用前面在模拟 B1 中列出的分布组，根据相同分布生成了相等数量的 k 样本。我们在设计中选择数量为 $k = 3$、$k = 4$ 和 $k = 6$ 的样本，并将每次试验中的 k 样本数量固定为 10, 15、20、25、50 和 100。

我们对每种设计情况的同一样本执行了目标显著性水平为 $\alpha = 0.05$ 的双侧 MC 检验。由于在每种情况下，模拟显著性水平基于 10,000 对样本仿行，并且我们使用目标显著性水平 5%，因此，模拟误差为 $\sqrt{0.95(0.05)/10{,}000} = 0.2\%$。

模拟结果汇总如下表 2a 和 2b 中所示。

表 2a 平衡多样本设计中双侧多重比较检验的模拟显著性水平。此检验的目标显著性水平为 0.05。

分布	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	模拟水平	n_i	模拟水平	n_i	模拟水平
N(0,1)	10	0.038	10	0.038	10	0.036
	15	0.040	15	0.041	15	0.039
	20	0.039	20	0.040	20	0.041
	25	0.045	25	0.047	25	0.047
	50	0.046	50	0.046	50	0.052
	100	0.049	100	0.049	100	0.052
t(5)	10	0.042	10	0.044	10	0.042
	15	0.041	15	0.044	15	0.046
	20	0.043	20	0.045	20	0.045
	25	0.046	25	0.048	25	0.046
	50	0.040	50	0.039	50	0.038
	100	0.038	100	0.040	100	0.040
T(10)	10	0.033	10	0.037	10	0.038

分布	$k=3$ $n_1=n_2=n_3$		$k=4$ $n_1=n_2=n_3=n_4$		$k=6$ $n_1=n_2=\cdots=n_6$	
	n_i	模拟水平	n_i	模拟水平	n_i	模拟水平
	15	0.040	15	0.042	15	0.041
	20	0.042	20	0.043	20	0.043
	25	0.041	25	0.042	25	0.045
	50	0.047	50	0.044	50	0.047
	100	0.048	100	0.046	100	0.047
Lpl	10	0.056	10	0.063	10	0.071
	15	0.056	15	0.061	15	0.063
	20	0.054	20	0.058	20	0.059
	25	0.051	25	0.056	25	0.58
	50	0.045	50	0.051	50	0.049
	100	0.044	100	0.047	100	0.050
B(3,3)	10	0.031	10	0.031	10	0.031
	15	0.037	15	0.036	15	0.034
	20	0.035	20	0.036	20	0.037
	25	0.039	25	0.038	25	0.040
	50	0.044	50	0.044	50	0.044
	100	0.044	100	0.046	100	0.043
U(0,1)	10	0.029	10	0.025	10	0.023
	15	0.026	15	0.027	15	0.026
	20	0.028	20	0.030	20	0.028
	25	0.034	25	0.033	25	0.032
	50	0.041	50	0.036	50	0.036
	100	0.048	100	0.047	100	0.045

分布	$k=3$ $n_1=n_2=n_3$		$k=4$ $n_1=n_2=n_3=n_4$		$k=6$ $n_1=n_2=\cdots=n_6$	
	n_i	模拟水平	n_i	模拟水平	n_i	模拟水平
Exp	10	0.063	10	0.073	10	0.076
	15	0.056	15	0.058	15	0.064
	20	0.051	20	0.053	20	0.057
	25	0.043	25	0.045	25	0.050
	50	0.033	50	0.037	50	0.038
	100	0.033	100	0.035	100	0.035

表 2b　平衡多样本设计中双侧多重比较检验的模拟显著性水平。此检验的目标显著性水平为 0.05。

分布	$k=3$ $n_1=n_2=n_3$		$k=4$ $n_1=n_2=n_3=n_4$		$k=6$ $n_1=n_2=\cdots=n_6$	
	n_i	模拟水平	n_i	模拟水平	n_i	模拟水平
Chi(5)	10	0.040	10	0.046	10	0.048
	15	0.043	15	0.046	15	0.049
	20	0.040	20	0.040	20	0.042
	25	0.040	25	0.045	25	0.042
	50	0.037	50	0.038	50	0.040
	100	0.036	100	0.037	100	0.038
Chi(10)	10	0.042	10	0.045	10	0.045
	15	0.038	15	0.044	15	0.047
	20	0.036	20	0.039	20	0.040
	25	0.043	25	0.044	25	0.045
	50	0.041	50	0.040	50	0.042
	100	0.038	100	0.040	100	0.042
B(8,1)	10	0.058	10	0.060	10	0.066
	15	0.057	15	0.061	15	0.064
	20	0.049	20	0.051	20	0.055

分布	$k=3$ $n_1=n_2=n_3$		$k=4$ $n_1=n_2=n_3=n_4$		$k=6$ $n_1=n_2=\cdots=n_6$	
	n_i	模拟水平	n_i	模拟水平	n_i	模拟水平
	25	0.044	25	0.046	25	0.050
	50	0.037	50	0.037	50	0.039
	100	0.037	100	0.038	100	0.039
CN(0.9,3)	10	0.020	10	0.018	10	0.016
	15	0.022	15	0.020	15	0.017
	20	0.014	20	0.012	20	0.008
	25	0.011	25	0.011	25	0.008
	50	0.009	50	0.007	50	0.006
	100	0.010	100	0.008	100	0.008
CN(0.8, 3)	10	0.017	10	0.015	10	0.011
	15	0.013	15	0.011	15	0.008
	20	0.012	20	0.012	20	0.009
	25	0.013	25	0.010	25	0.009
	50	0.011	50	0.011	50	0.009
	100	0.014	100	0.012	100	0.010

如表 2a 和 2b 中所示，对于平衡设计中的对称和近对称分布，当样本数量较小时，MC 检验通常比较保守。另一方面，对于从高度偏移的分布（如，指数和 beta(8, 1) 分布）获取的小样本，检验比较宽松。但是，随着样本数量的增大，模拟显著性水平会越接近目标显著性水平(0.05)。此外，样本数似乎对中等样本的检验性能没有很大的影响。但如果数据被异常值污染，则会对检验性能产生显著影响。如果数据中存在异常值，则检验会持续过度保守。

部分 II：不平衡设计

我们执行了模拟，以检查不平衡设计中 MC 检验的性能。我们使用前面在模拟 B1 中所述的分布组，根据相同分布生成了 3 样本。在第一组试验中，前两个样本的数量是 $n_1=n_2=10$，第三个样本的数量是 $n_3=15,20,25,50,100$。在第二组试验中，前两个样本的数量是 $n_1=n_2=15$，第三个样本的数量是 $n_3=20,25,30,50,100$。在第三组试验中，我们将最小样本数量设置为 20，前两个样本的数量是 $n_1=n_2=20$，第三个样本的数量是 $n_3=25,30,40,50,100$.

我们对每种分布中相同的三个样本执行了目标显著性水平为 $\alpha=0.05$ 的双侧 MC 检验。由于在每种情况下，模拟显著性水平基于 10,000 对样本仿行，并且我们使用目标显著性水平 5%，因此，模拟误差为 $\sqrt{0.95(0.05)/10{,}000}=0.2\%$.

模拟结果汇总如下表 3a 和 3b 中所示。

表 3a　多样本不平衡设计中多重比较检验的模拟显著性水平。此检验的目标显著性水平为 0.05。

分布	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	模拟水平	n_3	模拟水平	n_3	模拟水平
N(0,1)	15	0.032	20	0.040	25	0.045
	20	0.037	25	0.039	30	0.041
	25	0.038	30	0.037	40	0.043
	50	0.041	50	0.044	50	0.041
	100	0.042	100	0.042	100	0.044
t(5)	15	0.040	20	0.042	25	0.043
	20	0.036	25	0.040	30	0.037
	25	0.044	30	0.036	40	0.038
	50	0.033	50	0.036	50	0.035
	100	0.032	100	0.031	100	0.032
t(10)	15	0.039	20	0.042	25	0.042
	20	0.038	25	0.041	30	0.040
	25	0.040	30	0.041	40	0.041
	50	0.037	50	0.043	50	0.042
	100	0.036	100	0.039	100	0.040
Lpl	15	0.059	20	0.060	25	0.054
	20	0.057	25	0.054	30	0.051
	25	0.056	30	0.051	40	0.050
	50	0.049	50	0.051	50	0.050
	100	0.048	100	0.047	100	0.046
B(3,3)	15	0.034	20	0.033	25	0.037
	20	0.031	25	0.035	30	0.039
	25	0.031	30	0.034	40	0.039
	50	0.036	50	0.039	50	0.038

分布	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	模拟水平	n_3	模拟水平	n_3	模拟水平
	100	0.035	100	0.039	100	0.039
U(0,1)	15	0.027	20	0.030	25	0.032
	20	0.030	25	0.030	30	0.031
	25	0.028	30	0.032	40	0.036
	50	0.039	50	0.034	50	0.037
	100	0.042	100	0.038	100	0.042
Exp	15	0.061	20	0.053	25	0.042
	20	0.060	25	0.052	30	0.047
	25	0.054	30	0.049	40	0.043
	50	0.050	50	0.046	50	0.041
	100	0.044	100	0.040	100	0.040

表 3b　多样本不平衡设计中 MC 检验的模拟显著性水平。此检验的目标显著性水平为 0.05。

分布	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	模拟水平	n_3	模拟水平	n_3	模拟水平
Chi(5)	15	0.047	20	0.045	25	0.041
	20	0.043	25	0.042	30	0.039
	25	0.043	30	0.039	40	0.040
	50	0.039	50	0.037	50	0.040
	100	0.034	100	0.035	100	0.034
Chi(10)	15	0.043	20	0.042	25	0.042
	20	0.039	25	0.038	30	0.041
	25	0.040	30	0.041	40	0.038
	50	0.038	50	0.041	50	0.042
	100	0.035	100	0.034	100	0.035
B(8,1)	15	0.056	20	0.052	25	0.048

分布	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	模拟水平	n_3	模拟水平	n_3	模拟水平
	20	0.054	25	0.046	30	0.044
	25	0.050	30	0.047	40	0.046
	50	0.046	50	0.043	50	0.043
	100	0.043	100	0.042	100	0.044
CN(0.9,3)	15	0.017	20	0.020	25	0.017
	20	0.020	25	0.019	30	0.012
	25	0.017	30	0.016	40	0.013
	50	0.019	50	0.016	50	0.012
	100	0.014	100	0.016	100	0.010
CN(0.8, 3)	15	0.012	20	0.013	25	0.013
	20	0.016	25	0.012	30	0.012
	25	0.014	30	0.010	40	0.010
	50	0.015	50	0.010	50	0.013
	100	0.012	100	0.011	100	0.010

表 3a 和 3b 中所示的模拟显著性水平与之前针对平衡设计的多个样本报告的模拟显著性水平一致。因此，MC 检验的性能似乎不受不平衡设计影响。此外，如果最小样本数量至少为 20，则模拟显著性水平接近目标显著性水平，但受污染数据除外。

结论是，对于平衡和不平衡设计中的多个 (k) 样本，如果最小样本数量至少为 20，则 MC 检验的执行效果不错。但是，当样本较小时，对于对称和近对称数据，检验比较保守，对于高度偏斜的数据，检验比较宽松。

附录 C：理论功效函数

目前没有 MC 检验的确切理论功效函数。但是，对于双样本设计，可以获得基于大样本理论方法的近似功效函数。对于多样本设计，需要进一步的研究才能推导出类似的近似功效函数。

但是，对于双样本设计，可以使用大样本理论方法获得 Bonett 检验的理论功效函数。更具体地说，下面给定的检验统计量 T 按自由度为 1 的卡方分布进行异步分布：

$$T = \frac{(\ln \hat{\rho}^2 - \ln \rho^2)^2}{\frac{\gamma - g_1}{n_1 - 1} + \frac{\gamma - g_2}{n_2 - 1}}$$

在这种 T、$\hat{\rho} = S_1/S_2$、$\rho = \sigma_1/\sigma_2$、$g_i = (n_i - 3)/n_i$ 的表达式中，γ 是两个总体的未知常用峰度。

然后，它遵循以下原则：近似显著性水平为 α 的方差等同性的双侧 Bonett 检验的理论功效函数按如下给定

$$\pi(n_1, n_2, \rho) = 1 - \Phi\left(z_{\alpha/2} - \frac{\ln \rho^2}{se}\right) + \Phi\left(-z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

其中

$$se = \sqrt{\frac{\gamma - g_1}{n_1 - 1} + \frac{\gamma - g_2}{n_2 - 1}}$$

对于单侧检验，在检验 $\sigma_1 > \sigma_2$ 时，近似功效函数为

$$\pi(n_1, n_2, \rho) = 1 - \Phi\left(z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

在检验 $\sigma_1 < \sigma_2$ 时，近似功效函数为

$$\pi(n_1, n_2, \rho) = \Phi\left(-z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

请注意，在数据分析的样本数量计划阶段，总体的常用峰度 γ 未知。因此，调查人员通常必须根据专家意见或以前的试验结果才能获取 γ 的计划值。如果这些信息不可用，最好开展小规模初步研究，以制定主要研究计划。使用初步研究中的样本，获得的计划值 γ 将作为按如下给定的合并峰度

$$\hat{\gamma}_P = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(X_{1j} - m_1)^4 + \sum_{j=1}^{n_2}(X_{2j} - m_2)^4}{[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]^2}$$

在“协助”菜单中，将根据用户的现有数据按回溯方式获取 γ 的计划估计值。

附录 D：比较理论和模拟功效

模拟 D1：Bonett 检验的模拟（实际）功效

我们进行了模拟，以将 Bonett 检验的模拟功效水平与基于附录 C 中推导出的近似功效函数的功效水平进行比较。

我们为上述每种分布生成了 10,000 对样本（参见模拟 B1）。通常，根据模拟 B1 中以前的结果，选定的样本数量足够大，从而检验的模拟显著性水平与目标显著性水平相当接近。

为按标准差比率 $\rho = \sigma_1/\sigma_2 = 1/2$ 评估模拟功效水平，我们将每对样本中的第二个样本乘以常数 2。结果是，对于给定分布和给定样本数量 n_1 和 n_2，模拟功效水平计算为 10,000 对样本仿行（其双侧 Bonett 检验效果显著）的一部分。检验的目标显著性水平固定为 $\alpha = 0.05$。要进行比较，我们根据附录 C 中推导出的近似功效函数计算出了对应的理论功效水平。

结果如下表 4 中所示。

表 4　将双侧 Bonett 检验的模拟功效水平与近似功效水平进行比较。目标显著性水平为 0.05。

分布	n_1, n_2	近似功效	模拟功效	分布	n_1, n_2	近似功效	模拟功效
N(0,1)	20，10	0.627	0.527	Exp	20，10	0.222	0.227
	20，20	0.83	0.765		20，20	0.322	0.368
	20，30	0.896	0.846		20，30	0.377	0.434
	20，40	0.925	0.886		20，40	0.412	0.475
	30，15	0.825	0.771		30，15	0.32	0.307
	30，30	0.954	0.925		30，30	0.458	0.50
	30，45	0.98	0.97		30，45	0.531	0.579
	30，60	0.989	0.984		30，60	0.575	0.622
t(5)	20，10	0.222	0.379	Chi(5)	20，10	0.355	0.347
	20，20	0.322	0.569		20，20	0.517	0.53
	20，30	0.377	0.637		20，30	0.597	0.616
	20，40	0.412	0.69		20，40	0.644	0.661
	30，15	0.32	0.545		30，15	0.513	0.51
	30，30	0.458	0.733		30，30	0.701	0.711
	30，45	0.531	0.795		30，45	0.781	0.793

分布	n_1, n_2	近似功效	模拟功效	分布	n_1, n_2	近似功效	模拟功效
	30，60	0.575	0.828		30，60	0.823	0.833
t(10)	20，10	0.476	0.45	Chi(10)	20，10	0.454	0.414
	20，20	0.673	0.673		20，20	0.646	0.631
	20，30	0.756	0.749		20，30	0.73	0.717
	20，40	0.80	0.803		20，40	0.776	0.771
	30，15	0.668	0.659		30，15	0.641	0.618
	30，30	0.85	0.852		30，30	0.828	0.819
	30，45	0.91	0.911		30，45	0.892	0.882
	30，60	0.936	0.937		30，60	0.921	0.912
Lpl	20，10	0.321	0.33	B(8,1)	20，10	0.363	0.278
	20，20	0.469	0.519		20，20	0.528	0.463
	20，30	0.545	0.585		20，30	0.609	0.549
	20，40	0.59	0.632		20，40	0.655	0.60
	30，15	0.466	0.475		30，15	0.524	0.419
	30，30	0.647	0.673		30，30	0.713	0.634
	30，45	0.729	0.758		30，45	0.792	0.737
	30，60	0.773	0.80		30，60	0.833	0.777
B(3,3)	20，10	0.777	0.628	CN(0.9,3)	20，10	0.238	0.284
	20，20	0.939	0.869		20，20	0.346	0.452
	20，30	0.973	0.936		20，30	0.405	0.517
	20，40	0.984	0.964		20，40	0.442	0.561
	30，15	0.935	0.871		30，15	0.343	0.374
	30，30	0.993	0.98		30，30	0.491	0.598
	30，45	0.998	0.995		30，45	0.567	0.70
	30，60	0.999	0.999		30，60	0.612	0.719
U(0,1)	20，10	0.916	0.74	CN(0.8,3)	20，10	0.26	0.223

分布	n_1, n_2	近似功效	模拟功效	分布	n_1, n_2	近似功效	模拟功效
	20，20	0.992	0.95		20，20	0.379	0.396
	20，30	0.998	0.985		20，30	0.444	0.467
	20，40	0.999	0.995		20，40	0.484	0.52
	30，15	0.991	0.941		30，15	0.376	0.354
	30，30	1.0	0.996		30，30	0.535	0.549
	30，45	1.0	1.0		30，45	0.614	0.65
	30，60	1.0	1.0		30，60	0.661	0.706

结果显示，通常，近似功效水平与模拟功效水平互相接近。它们会随着样本数量增大而越来越接近。对于轻尾到中尾的对称和近对称分布，近似功效水平通常略高于模拟功效水平。但是，对于重尾对称分布或高度偏斜分布，近似功效水平略低于模拟功效水平。这两个功效函数之间的差异通常并不重要，但根据自由度为 5 的 t 分布生成样本的情况除外。

总之，当最小样本数量达到 20 时，近似功效水平与模拟功效水平极为接近。因此，样本数量计划可基于近似功效函数进行。