单因子方差分析

概述

单因子方差分析用于比较三组或更多组的平均值，以确定它们之间是否存在显著的不同。另一个重要功能是估计特定组之间的差值。

单因子方差分析中最常用的组差异检验方法是 F 检验，它基于一个假设，即所有样本的总体具有一个共同而且未知的标准差。我们认识到，在实际应用中样本往往具有不同的标准差。因此，我们想要研究 Welch 方法，它作为 F 检验的替代方法，可以处理不同的标准差。我们还想要制定一种方法，用于计算对具有不同标准差的样本进行说明的多重比较。采用这种方法，我们可以绘制各个区间，从而提供一种简单的方式来识别各组之间的差异。

在本文中，我们将介绍如何制定在 Minitab 协助单因子方差分析过程中使用的方法，用于：

- Welch 检验
- 多重比较区间

此外，我们还研究影响单因子方差分析结果的有效性的条件，包括异常数据的状态、样本量和检验功效以及数据的正态性。根据这些条件，“协助”会自动对您的数据进行以下检查并在“报告卡”中报告研究结果：

- 异常数据
- 样本量
- 数据的正态性

在本文中，我们探讨了在实际应用中这些条件与单因子方差分析之间的关系，还介绍了如何制定相关指南来在“协助”中检查这些条件。

单因子方差分析方法

F 检验与 Welch 检验

单因子方差分析中常用的 F 检验基于一个假设，即所有分组具有一个共同而且未知的标准差（σ）。在实际应用中，此假设很少成立，这会引起对 I 类错误发生概率的控制问题。I 类错误是指错误地否定原假设的概率（样本不存在差异时误断为存在显著差异）。当样本的不同分组具有不同的标准差时，检验得出错误结论的可能性会更大。为了解决此问题，作为 F 检验替代方法的 Welch 检验应运而生 (Welch, 1951)。

目标

我们需要确定要在“协助”的单因子方差分析过程中使用 F 检验还是 Welch 检验。为此，我们需要评估 F 检验和 Welch 检验的实际检验结果与检验的目标显著性水平（alpha 值或 I 类错误概率）的匹配程度；即在不同样本量和样本标准差的条件下，观察假设检验错误地否定原假设的频率是高于还是低于所设目标。

方法

为了比较 F 检验和 Welch 检验，我们通过改变样本数目、样本量和样本标准差的方式进行了多次模拟。在每个条件下，我们分别使用 F 检验和 Welch 方法进行了 10,000 次方差分析。我们生成随机数据，并使样本的平均值相同，这样对于每次检验，其原假设均成立。然后，我们分别使用 0.05 和 0.01 的目标显著性水平进行了检验。我们分别计算了这 10,000 次检验中 F 检验和 Welch 检验实际否定原假设的次数，并将此比率与目标显著性水平进行了比较。如果检验方法很有效，则检验的 I 类错误概率应非常接近目标显著性水平。

结果

我们发现在检验的所有条件下，使用 Welch 方法的结果与 F 检验的结果相当甚至更好。例如，使用 Welch 检验比较 5 个样本时，I 类错误概率在 0.0460 和 0.0540 之间，非常接近于 0.05 的目标显著性水平。这表明即使样本量与标准差因样本不同而不同，Welch 方法的 I 类错误概率也与目标值相近。

在另一方面，F 检验的 I 类错误概率则介于 0.0273 和 0.2277 之间。在下列条件下，F 检验方法效果不佳：

- 当最大样本的标准差也最大时，I 类错误概率低于 0.05。这种情况会导致检验结果更保守，同时表明，当样本的标准差不同时，仅增加样本量不是一个可行的解决方案。
- 当样本量相同，但标准差不同时，I 类错误概率为 0.05 以上。当标准差较大的样本量比其他样本小时，该比率也大于 0.05。尤其是，当较小的样本具有较大的标准差时，该检验错误地否定原假设的风险会大大增加。

有关模拟方法和结果的详细信息，请参见附录 A。

当样本的标准差和大小不同时，Welch 方法比较有效，因此我们在“协助”中采用 Welch 方法来执行单因子方差分析过程。

比较区间

当方差分析具有重要的统计意义时，表明至少有一个样本平均值不同于其他平均值，分析的下一步就是确定哪些样本具有重要的统计意义。进行这种比较的直观方式是绘制置信区间图并确定区间不重叠的样本。但是，各个置信区间不专门用于比较，因此从图中得出的结论可能与检验结果不匹配。虽然已为标准差相同的样本发布了多重比较方法，不过我们仍需要扩展这种方法，以考虑标准差不同的样本。

目标

我们希望制定一种方法来计算可在样本之间进行比较并且尽可能匹配检验结果的各个比较区间。我们还想提供一种直观的方法，用于确定哪些样本的统计方式不同于其他样本。

方法

标准的多重比较方法 (Hsu 1996) 提供了每对平均值之间的差异区间，同时控制在进行多重比较时增加的错误。在样本量相同以及假设标准差相同的特殊情况下，也可以通过与所有平均值对的差异区间完全对应的方式显示每个平均值的各个区间。在样本量不同以及假设标准差相同的情况下，Hochberg, Weiss, and Hart (1982) 根据 Tukey-Kramer 多重比较方法，制定了各个区间，大致相当于平均值对之间的差异区间。在“协助”中，我们将采用与 Games-Howell 多重比较方法相同的方法，但假设标准差不同。Minitab 16 的“协助”中使用的方法概念相仿，但不以 Games-Howell 方法为主。有关详细信息，请参见附录 B。

结果

“协助”在单因子方差分析总结报告的平均值比较表中显示比较区间。当方差分析具有重要的统计意义时，不与至少一个其他区间重叠的任何比较区间将被标记为红色。检验和比较区间有可能不符，不过这种结果比较少见，因为这两种方法否定原假设成立的概率相同。如果方差分析具有重要意义，但是所有区间都重叠，那么重叠最少的平均值对将被标记为红色。如果方差分析不具有重要的统计意义，则任何区间都不会被标记为红色，即使某些区间不重叠也如此。

数据检查

异常数据

异常数据是非常大或非常小的数据值，也称为异常值。异常数据会对分析结果产生巨大的影响，并且可能影响发现具有重要统计意义的结果的概率，特别是当样本较小时。异常数据可以表明数据收集问题，或者由您正在研究的过程的异常表现产生的问题。因此，这些数据点往往值得研究，应尽可能予以更正。

目标

我们想要制定一种方法来检查相对于总体样本而言，非常大或非常小的数据值，这可能会影响分析的结果。

方法

我们制定了一种方法，用于根据 Hoaglin, Iglewicz, and Tukey (1986) 所述的方法检查异常数据，以确定箱线图中的异常值。

结果

如果某个数据点超出分布范围下限或上限 1.5 倍的四分位范围，“协助”将该数据点识别为异常数据点。上、下四分位数分别是数据的第 25 个和第 75 个百分位数。四分位范围是两个四分位数之间的差异。即使有多个异常值，这种方法也能正常使用，因为它可以检测到每一个具体的异常值。

当检查异常数据时，“协助”会在报告卡中显示以下状态指标：

状态	条件
	没有异常数据点。
	至少有一个异常数据点，可能会对结果产生巨大的影响。

样本量

功效是任何假设检验的重要特性，因为它指示当存在显著影响或差异时找到这些影响或差异的可能性。功效是否定原假设，赞成另一种假设的可能性。提高检验功效最简便的方法通常是增加样本量。在“协助”中，对于低功效检验，我们指明您需要多大的样本才能发现指定的差异。如果没有指定差异，我们将指明您需要用足够的功效才能发现的差异。要提供这些信息，我们需要制定一种计算功效的方法，因为“协助”使用的 Welch 方法中没有准确的功效公式。

目标

要制定计算功效的方法，我们需要解决两个问题。首先，“协助”不要求用户输入一组完整的平均值；它仅要求他们输入具有实际意义的平均值之间的差异。对于任何给定差异，可以无限

配置产生差异的平均值。因此，我们需要制定一个合理的方法，以确定使用哪个平均值来计算功效，因为我们无法为所有可能的平均值配置计算功效。其次，我们需要制定一种方法来计算功效，因为“协助”使用 Welch 方法，该方法不需要相同的样本量或标准差。

方法

为了解决无限配置平均值问题，我们根据 Minitab（统计 >方差分析 > 单因子）中的标准单因子方差分析过程中使用的方法制定了一种方法。我们关注只有两个平均值的规定数量不同，而其他平均值相等（设为平均值的加权平均值）的情况。因为我们假设只有两个平均值与其他所有平均值（不超过两个）不同，所以该方法提供保守的功效估计。但是，因为样本可能具有不同的大小或标准差，功效计算仍然取决于假定哪两种方法不同。

为了解决这个问题，我们将确定两对平均值，分别代表最好和最坏的情况。最坏的情况发生在样本量相对样本方差较小时，此时功效最小；最好的情况发生在样本量相对样本方差较大时，此时功效最大。所有功效计算都考虑以下两种极端情况，假设确实有两个平均值与整个加权平均值不同，功效要么最小，要么最大。

为了制定功效计算方法，我们使用了 Kulinskaya et al. (2003) 中显示的一种方法。我们通过模拟（即我们为解决平均值配置问题制定的方法）比较了功效计算结果，该方法显示在 Kulinskaya et al. (2003) 中。我们还查看了另一个功效近似值，它更清晰地显示了功效与平均值配置之间的关系。有关功效计算的详细信息，请参见附录 C。

结果

这些方法的比较结果表明，Kulinskaya 方法提供较好的功效近似值，而我们处理平均值配置的方法也是合适的。

当数据未提供足够的原假设证据时，“协助”将计算用给定样本量 80% 和 90% 的概率检测到的实际差异。另外，如果指定实际差异，“协助”将计算此差异的最小和最大功效值。当功效值低于 90% 时，“协助”将根据指定差值和观测到的样本标准差计算样本量。为了确保样本量实现的最小功效值和最大功效值在 90% 或以上，我们假设指定的差值位于可变性最大的两个平均值之间。

如果用户没有指定差值，“协助”会发现最大的差值，其中功效值的最大范围为 60%。此值标记在功效报告上的红条和黄条之间的边界处，对应 60% 的功效。我们还会发现功效值的最小范围是 90% 的最小差值。此值标记在功效报告上的红条和黄条之间的边界处，对应 90% 的功效。

检查功效和样本量时，“协助”在报告卡中显示以下状态指标：

状态	条件
	数据未提供足够的证据，以得出平均值之间存在差异的结论。未指定差异。
	检验发现平均值之间存在差异，所以功效不是问题。 或 功效是足够的。检验未发现平均值之间存在差异，但样本量足够大，至少有 90% 的机会检测到给定差异。
	功效可能足够。检验未发现平均值之间存在差异，但样品量足够大，有 80%～90% 的机会检测到给定差异。报告实现 90% 的功效所需的样本量。

状态	条件
	功效可能不够。检验未发现平均值之间存在差异，但样本量足够大，有 60%～80% 的机会检测到给定差异。报告实现 80% 和 90% 的功效所需的样本量。
	功效不够。检验未发现平均值之间存在差异，样本量不够大，不足以提供至少 60% 的机会检测到给定差异。报告实现 80% 和 90% 的功效所需的样本量。

正态性

许多统计方法中的一个共同假设是数据呈正态分布。幸运的是，即使数据不呈正态分布，基于正态假设的方法也同样有效。这在中心极限定理中有部分说明，其中提到，任何样本平均值都呈近似正态分布，当样本量变大时，近似值会更接近正态分布。

目标

我们的目标是确定需要多大的样本才能提供一个相当不错的正态分布近似值。我们想要通过呈各种非正态分布的小到中等样本研究 Welch 检验和比较区间。我们需要确定 Welch 方法和比较区间的实际检验结果与检验选定的显著性水平（alpha 值或 I 类错误概率）的匹配程度；即在不同样本量、级数和非正态分布的条件下，观察假设检验错误地否定原假设的频率是高于还是低于设定目标。

方法

为了估计 I 类错误概率，我们通过改变样本数目、样本量和数据分布的方式进行了多次模拟。模拟包括与正态分布显著不同的偏态分布和重尾分布。在各种检验中，多个样本的大小和标准差恒定不变。

在各种条件下，我们分别使用 Welch 方法和比较区间进行了 10,000 次方差分析。我们生成随机数据，并使样本的平均值相同，这样对于每次检验，其原假设均成立。然后，我们使用 0.05 的目标显著性水平进行了检验。我们计算了这 10,000 次检验中实际否定原假设的次数，并将此比率与目标显著性水平进行了比较。对于比较区间，我们计算了这 10,000 次检验中区间表示一个或多个差异的次数。如果检验方法很有效，则 I 类错误概率应非常接近目标显著性水平。

结果

总体上，样本量小至 10 或 15 的检验和比较区间在所有条件下都很有效。对于 9 个或更少的样本检验，在几乎所有情况下，10 个样本量的检验结果都在目标显著性水平的 3 个百分点范围内，15 个样本量的检验结果则在 2 个百分点范围内。对于 10 个或更多的样本检验，在大多数情况下，对于 15 个样本，检验结果在 3 个百分点范围内，对于 20 个样本，则在 2 个百分点范围内。有关详细信息，请参见附录 D。

因为相对较小的样本检验很有效，“协助”没有检验数据的正态性。不过，“协助”会检查样本量并指示 2-9 级的样本少于 15 个，10-12 级的样本少于 20 个。根据这些结果，“协助”在报告卡中显示以下状态指标：

状态	条件
	样本量至少为 15 或 20，所以正态性不是问题。
	某些样本量小于 15 或 20，因此可能存在正态性问题。

参考书

Dunnet, C. W. (1980). Pairwise Multiple Comparisons in the Unequal Variance Case. *Journal of the American Statistical Association, 75, 796-800.*

Hoaglin, D. C., Iglewicz, B., and Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association, 81, 991-999.*

Hochberg, Y., Weiss G., and Hart, S. (1982). On graphical procedures for multiple comparisons. *Journal of the American Statistical Association, 77, 767-772.*

Hsu, J. (1996). *Multiple comparisons: Theory and methods*. Boca Raton, FL: Chapman & Hall.

Kulinskaya, E., Staudte, R. G., and Gao, H. (2003). Power approximations in testing for unequal means in a One-Way ANOVA weighted for unequal variances, *Communication in Statistics, 32 (12), 2353-2371.*

Welch, B.L. (1947). The generalization of “Student’s” problem when several different population variances are involved. *Biometrika, 34, 28-35*

Welch, B.L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika 38, 330-336.*

附录 A：F 检验与 Welch 检验

F 检验将在违反标准差相同假设条件时导致 I 类错误概率增加；Welch 检验旨在避免这些问题。

Welch 检验

观察到 k 总体中大小为 n_1，…，n_k 的随机样本。用 $\mu_1,\cdots,\mu_k$ 表示总体平均值，$\sigma_1^2,\ldots,\sigma_k^2$ 表示总体方差。用 $\bar{x}_1,\ldots,\bar{x}_k$ 表示样本平均值，$s_1^2,\ldots,s_k^2$ 表示样本方差。我们感兴趣的是假设检验：

H_0：$\mu_1 = \mu_2 = \cdots = \mu_k$

H_1：$\mu_i \neq \mu_j$ 表示 i，j 等。

Welch 检验用于检验 k 平均值与

$$W^* = \frac{\sum_{j=1}^{k} w_j(\bar{x}_j - \hat{\mu})^2/(k-1)}{1+[2(k-2)/(k^2-1)]\sum_{j=1}^{k} h_j}$$

F(k － 1，f) 分布的统计值是否相等，其中

$w_j = \frac{n_j}{s_j^2}$、

$W = \sum_{j=1}^{k} w_j$、

$\hat{\mu} = \frac{\sum_{j=1}^{k} w_j \bar{x}_j}{W}$、

$h_j = \frac{(1-w_j/W)^2}{n_j-1}$ 和

$f = \frac{k^2-1}{3\sum_{j=1}^{k} h_j}$。

如果 $W^* \geq F_{k-1,f,1-\alpha}$（F 分布的百分位）超过概率 α，则 Welch 检验否定原假设。

不同的标准差

在本节中，我们将证明 F 检验对标准差相同假设违例的敏感性，并将其与 Welch 检验进行比较。

下面是使用 5 个 N(0，σ2) 样本的单因子方差分析的结果。每一行都基于使用 F 检验和 Welch 检验的 10,000 次模拟。我们通过增加第五个样本的标准差，使之成为其他样本的两倍和四倍，来检验标准差的两个条件。我们检验了样本量的三个不同的条件：样本量相同、第五个样本大于其他样本、第五个样本小于其他样本。

表 1　模拟 F 检验和 Welch 检验的 I 类错误概率，其中 5 个样本的目标显著性水平 α = 0.05

标准差 (σ1, σ2, σ3, σ4, σ5)	样本量 (n1, n2, n3, n4, n5)	F 检验	Welch 检验
1, 1, 1, 1, 2	10, 10, 10, 10, 20	.0273	.0524
1, 1, 1, 1, 2	20, 20, 20, 20, 20	.0678	.0462
1, 1, 1, 1, 2	20, 20, 20, 20, 10	.1258	.0540
1, 1, 1, 1, 4	10, 10, 10, 10, 20	.0312	.0460
1, 1, 1, 1, 4	20, 20, 20, 20, 20	.1065	.0533
1, 1, 1, 1, 4	20, 20, 20, 20, 10	.2277	.0503

当样本量相同（第 2 行和第 5 行）时，F 检验错误地否定原假设的概率大于目标概率 0.05，当标准差之间的差距增大时，概率也会增大。减小标准差最大的样本量会让该问题变得更加棘手。另一方面，增加标准差最大的样本量会减小否定概率。然而，样本量增加得太多，否定概率会变得太小，不仅会让原假设条件下的检验比预期更加保守，还会对另一种假设条件下的检验的功效产生不利的影响。将这些结果与 Welch 检验比较，在任何情况下哪个值与 0.05 的目标显著性水平相吻合。

接下来，我们对 k=7 的样本进行了模拟。表中的每一行将总结 10,000 次模拟 F 检验。我们改变了样本的标准差和大小。目标显著性水平分别为 $\alpha = 0.05$ 和 $\alpha = 0.01$。正如上文所述，我们看到，目标值的偏差会相当严重。可变性较高时使用较小的样本量，将导致非常大的 I 类错误概率，使用更大的样本可能会导致非常保守的检验结果。结果在下面的表 2 中显示。

表 2　7 个样本的模拟 F 检验的 I 类错误概率

标准差 (σ1, σ2, σ3, σ4, σ5, σ6, σ7)	样本量 (n1, n2, n3, n4, n5, n6, n7)	目标 α = 0.05	目标 α = 0.01
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	21, 21, 21, 21, 22, 22, 12	0.0795	0.0233
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	20, 21, 21, 21, 21, 24, 12	0.0785	0.0226
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	20, 21, 21, 21, 21, 21, 15	0.0712	0.0199
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	20, 20, 20, 21, 21, 23, 15	0.0719	0.0172
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	20, 20, 20, 20, 21, 21, 18	0.0632	0.0166
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	20, 20, 20, 20, 20, 20, 20	0.0576	0.0138

标准差 (σ1, σ2, σ3, σ4, σ5, σ6, σ7)	样本量 (n1, n2, n3, n4, n5, n6, n7)	目标 α = 0.05	目标 α = 0.01
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	18, 19, 19, 20, 20, 20, 24	0.0474	0.0133
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	18, 18, 18, 18, 18, 18, 32	0.0314	0.0057
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	15, 18, 18, 19, 20, 20, 30	0.0400	0.0085
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	12, 18, 18, 18, 19, 19, 36	0.0288	0.0064
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	15, 15, 15, 15, 15, 15, 50	0.0163	0.0025
1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9	12, 12, 12, 12, 12, 12, 68	0.0052	0.0002
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	21, 21, 21, 21, 22, 22, 12	0.1097	0.0436
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	20, 21, 21, 21, 21, 24, 12	0.1119	0.0452
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	20, 21, 21, 21, 21, 21, 15	0.0996	0.0376
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	20, 20, 20, 21, 21, 23, 15	0.0657	0.0345
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	20, 20, 20, 20, 21, 21, 18	0.0779	0.0283
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	20, 20, 20, 20, 20, 20, 20	0.0737	0.0264
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	18, 19, 19, 20, 20, 20, 24	0.0604	0.0204
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	18, 18, 18, 18, 18, 18, 32	0.0368	0.0122
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	15, 18, 18, 19, 20, 20, 30	0.0390	0.0117
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	12, 18, 18, 18, 19, 19, 36	0.0232	0.0046
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	15, 15, 15, 15, 15, 15, 50	0.0124	0.0026
1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5	12, 12, 12, 12, 12, 12, 68	0.0027	0.0004

标准差 (σ1, σ2, σ3, σ4, σ5, σ6, σ7)	样本量 (n1, n2, n3, n4, n5, n6, n7)	目标 α = 0.05	目标 α = 0.01
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	21, 21, 21, 21, 22, 22, 12	0.134	0.0630
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	20, 21, 21, 21, 21, 24, 12	0.1329	0.0654
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	20, 21, 21, 21, 21, 21, 15	0.1101	0.0484
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	20, 20, 20, 21, 21, 23, 15	0.1121	0.0495
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	20, 20, 20, 20, 21, 21, 18	0.0876	0.0374
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	20, 20, 20, 20, 20, 20, 20	0.0808	0.0317
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	18, 19, 19, 20, 20, 20, 24	0.0606	0.0243
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	18, 18, 18, 18, 18, 18, 32	0.0356	0.0119
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	15, 18, 18, 19, 20, 20, 30	0.0412	0.0134
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	12, 18, 18, 18, 19, 19, 36	0.0261	0.0068
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	15, 15, 15, 15, 15, 15, 50	0.0100	0.0023
1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9	12, 12, 12, 12, 12, 12, 68	0.0017	0.0003
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	21, 21, 21, 21, 22, 22, 12	0.1773	0.1006
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	20, 21, 21, 21, 21, 24, 12	0.1811	0.1040
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	20, 21, 21, 21, 21, 21, 15	0.1445	0.0760
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	20, 20, 20, 21, 21, 23, 15	0.1448	0.0786
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	20, 20, 20, 20, 21, 21, 18	0.1164	0.0572
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	20, 20, 20, 20, 20, 20, 20	0.1020	0.0503

标准差 (σ1, σ2, σ3, σ4, σ5, σ6, σ7)	样本量 (n1, n2, n3, n4, n5, n6, n7)	目标 α = 0.05	目标 α = 0.01
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	18, 19, 19, 20, 20, 20, 24	0.0834	0.0369
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	18, 18, 18, 18, 18, 18, 32	0.0425	0.0159
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	15, 18, 18, 19, 20, 20, 30	0.0463	0.0168
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	12, 18, 18, 18, 19, 19, 36	0.0305	0.0103
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	15, 15, 15, 15, 15, 15, 50	0.0082	0.0021
1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7	12, 12, 12, 12, 12, 12, 68	0.0013	0.0001

附录 B：比较区间

平均值比较图允许您评估总体平均值之间具有重要统计意义的差异。

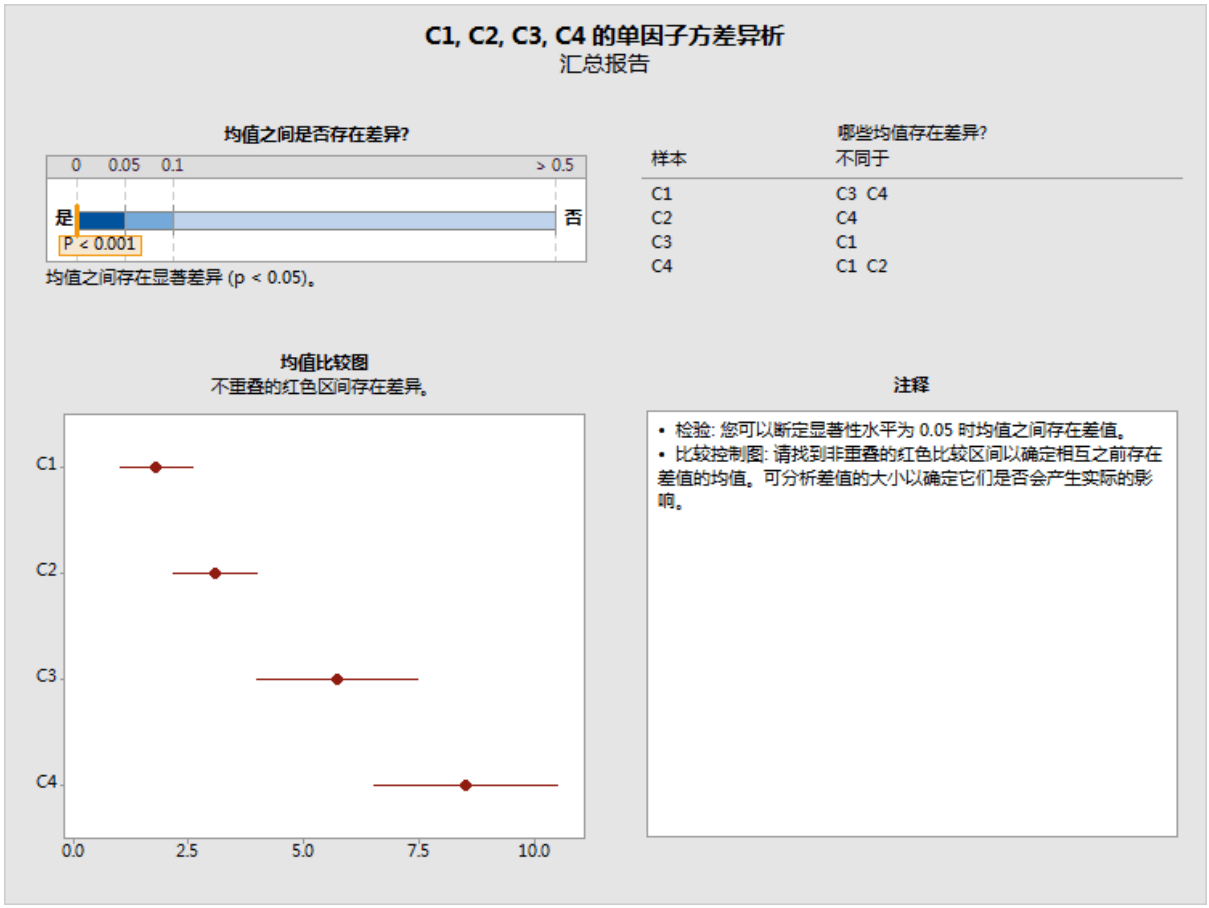
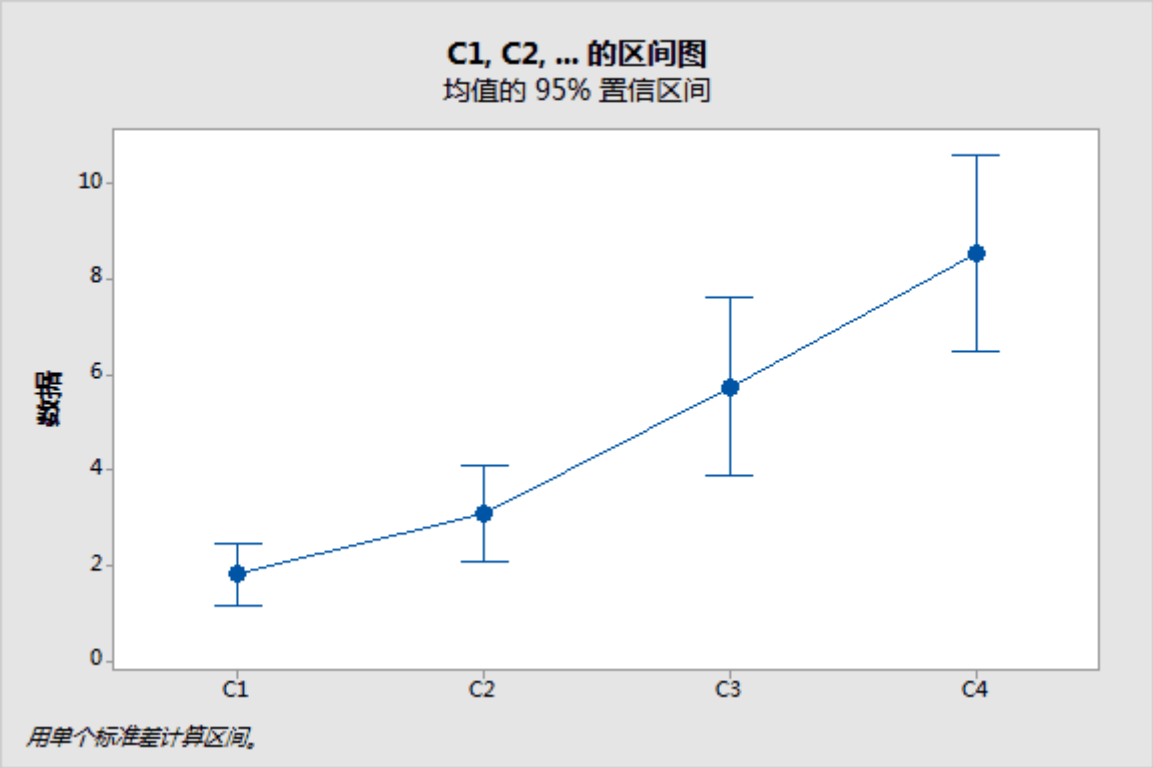


图 1 “协助”的单因子方差分析总结报告中的平均值比较图

在 Minitab（统计 > 方差分析 > 单因子）中的标准单因子方差分析过程的输出内容中显示的一组类似的区间：



不过，请注意，上面的区间仅仅是各个平均值的置信区间。当方差分析（F 或 Welch）得出某些平均值不同的结论时，自然就会查找不重叠的区间，并得出有关哪个平均值不同的结论。这种对各个置信区间的非正式分析常常会得出合理的结论，但不会以与方差分析相同的方式控制错误概率。根据总体数目，区间基本上可能会或多或少于检验结果，以得出存在差异的结论。因此，这两种方法可以轻松得出不一致的结论。进行多重比较时，比较图将与 Welch 检验结果更加一致，尽管它并不总是能够实现完全的一致性。

多重比较方法，如 Minitab（统计 >方差分析 > 单因子）中的 Tukey-Kramer 和 Games-Howell 比较方法，允许您绘制有关各个平均值之间的差异的具有统计意义的有效结论。这两种方法都是成对比较方法，其中提供每对平均值之间的差异的区间。所有区间同时包含所估计的差异的概率至少是 $1-\alpha$。Tukey-Kramer 方法取决于方差相同假设，而 Games-Howell 方法不需要方差相同。如果相等平均值的原假设成立，那么所有的差值都为零，而且任何 Games-Howell 区间不包含零的概率至多为 α。因此，我们可以使用区间执行显著性水平为 α 的假设检验。我们采用 Games-Howell 区间作为获得“协助”中的比较图区间的起点。

假设所有的 $\mu_i - \mu_j$, $1 \leq i < j \leq k$ 差异都有一组区间 $[L_{ij}, U_{ij}]$，我们希望为传达相同信息的各个平均值 μ_i, $1 \leq i \leq k$ 找到一组区间 $[L_i, U_i]$。这要求任何差值 d 都在区间 $[L_{ij}, U_{ij}]$ 内，除非存在 $\mu_i \in [L_i, U_i]$ 和 $\mu_j \in [L_j, U_j]$，使得 $\mu_i - \mu_j = d$.。区间的两端必须用等式关联起来。

$$U_i - L_j = U_{ij} \text{ 和}$$

$$L_i - U_j = L_{ij}.$$

对于 k = 2，我们只有一个差值，但有两个独立的区间，因此能够得到准确的比较区间。事实上，满足这一条件的区间宽度具有相当大的灵活性。对于 k = 3，有三个差值和三个独立的区

间，所以又能满足条件了，但是现在还不能灵活设定区间宽度。对于 k = 4，有六个差值，但只有四个独立的区间。比较区间必须设法使用较少的区间传达相同的信息。通常，对于 k ≥ 4，差值比平均值更多，因此没有确切的解决方案，除非将附加条件（如等宽）强加给差值区间。

仅当所有样本量都相同时，Tukey-Kramer 区间才具有相等的宽度。等宽也是假设方差相等的结果。Games-Howell 区间不假设方差相等，因此也不等宽。在“协助”中，我们将不得不借助近似方法来定义比较区间。

$\mu_i - \mu_j$ 的 Games-Howell 的区间是

$$\bar{x}_i - \bar{x}_j \pm \left|q^*(k,\hat{v}_{ij})\right|\sqrt{s_i^2/n_i + s_j^2/n_j}$$

其中 $q^*(k,\hat{v}_{ij})$ 是学生化范围分布相应的百分位数，取决于要比较的平均值数目 k，以及与平均值对 (i, j) 关联的自由度 v_{ij}。

$$\hat{v}_{ij} = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)^2}{\left(\frac{s_i^2}{n_i}\right)^2 \frac{1}{n_i - 1} + \left(\frac{s_j^2}{n_j}\right)^2 \frac{1}{n_j - 1}}.$$

采用以下方法，Hochberg, Weiss, and Hart (1982) 获得了大致相当于这些两两比较方法的独立区间：

$\bar{x}_i \pm |q^*(k,v)|s_p X_i$。

选择值 X_i 以尽量减少

$$\sum\sum_{i \neq j}\left(X_i + X_j - a_{ij}\right)^2,$$

其中：

$a_{ij} = \sqrt{1/n_i + 1/n_j}$。

我们通过从

$\bar{x}_i \pm d_i$ 表的 Games-Howell 比较结果中获得区间，来调整这种方法，以适应方差不同的情况。

选择值 d_i 以尽量减少

$$\sum\sum_{i \neq j}\left(d_i + d_j - b_{ij}\right)^2,$$

其中：

$b_{ij} = \left|q^*(k,\hat{v}_{ij})\right|\sqrt{s_i^2/n_i + s_j^2/n_j}$。

解决方案为

$d_i = \frac{1}{k-1}\sum_{j \neq i} b_{ij} - \frac{1}{(k-1)(k-2)}\sum_{j \neq i, l \neq i, j<l} b_{jl}$。

下图将使用以下两种方法比较 Welch 检验的模拟结果与比较区间的结果：一种是我们现在使用的 Games-Howell 方法，另一种是基于平均自由度的 Minitab 16 版本中使用的方法。纵轴是超过 10,000 次 Welch 检验错误地否定原假设或者部分比较区间重叠的模拟次数比例。这

些示例中的目标 α 值是 $\alpha = 0.05$。这些模拟涵盖各种标准差和样本量不同的情况；沿着水平轴的各个位置分别代表不同的情况。

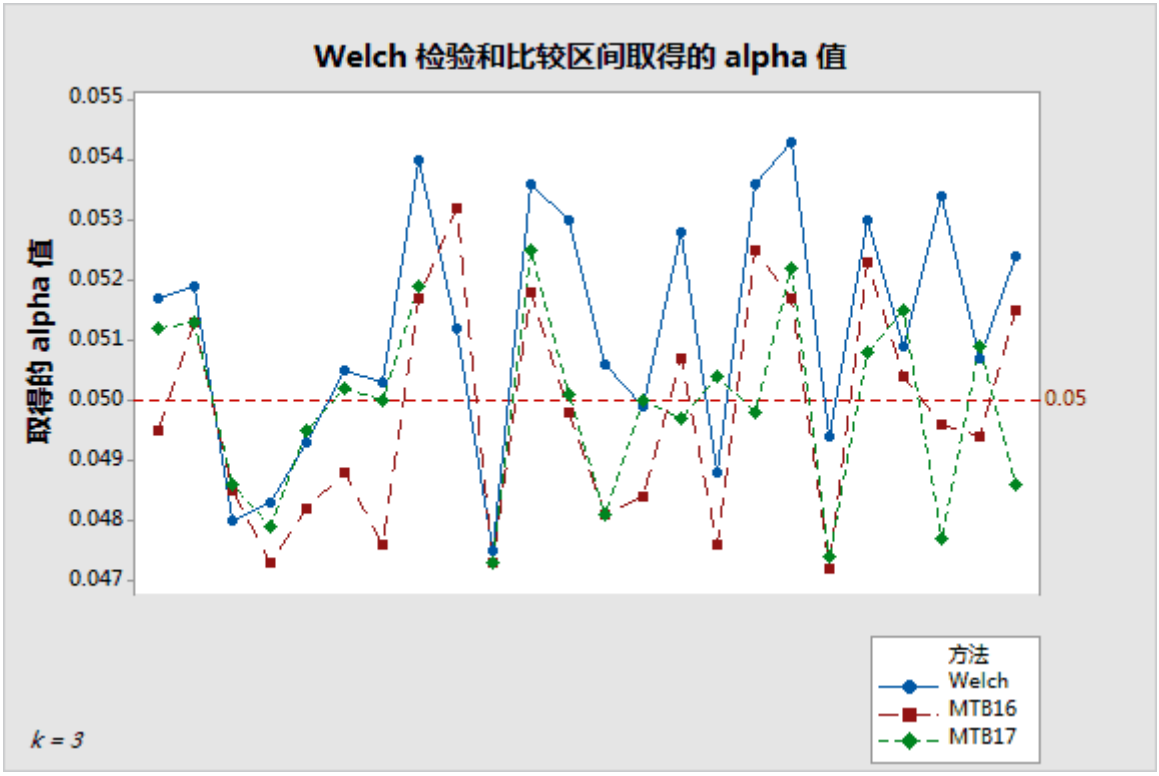


图 2　与计算 3 个样本的比较区间的两种方法比较的 Welch 检验

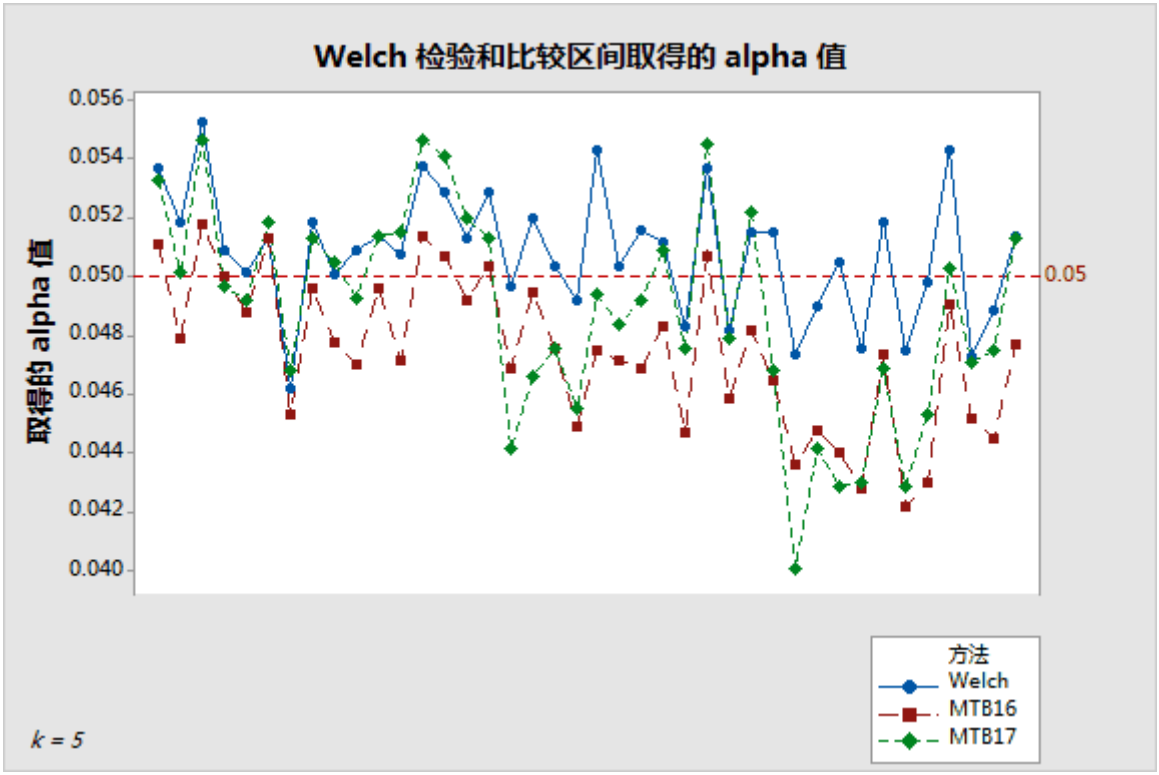


图 3　与计算 5 个样本的比较区间的两种方法比较的 Welch 检验

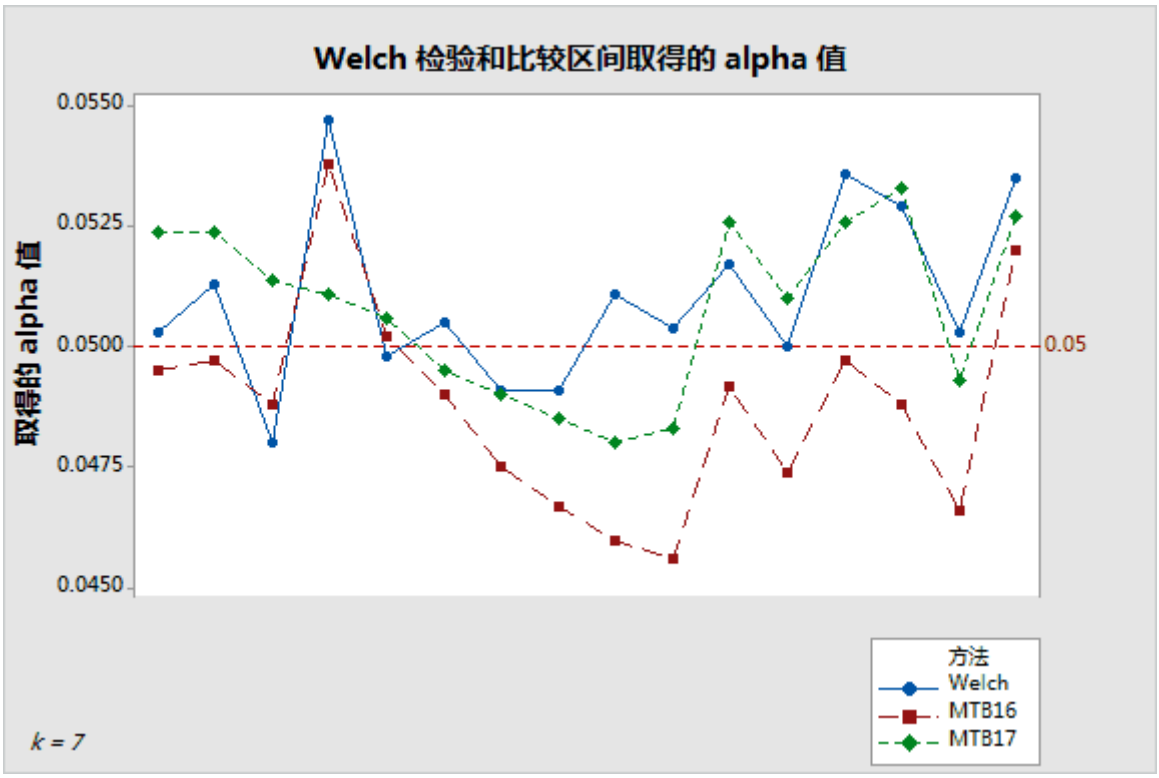


图 4　与计算 7 个样本的比较区间的两种方法比较的 Welch 检验

这些结果显示在围绕目标值 0.05 的狭窄范围内的模拟 alpha 值。此外，使用 Minitab 17 版本中实现的基于 Games-Howell 的方法的结果，与 Minitab 16 版本中使用的方法相比，大致更接近 Welch 检验的结果。

有证据表明，区间的覆盖概率可能与不同的标准差有关。但是敏感度不如 F 检验。下图说明了 K=5 时的这种关系。

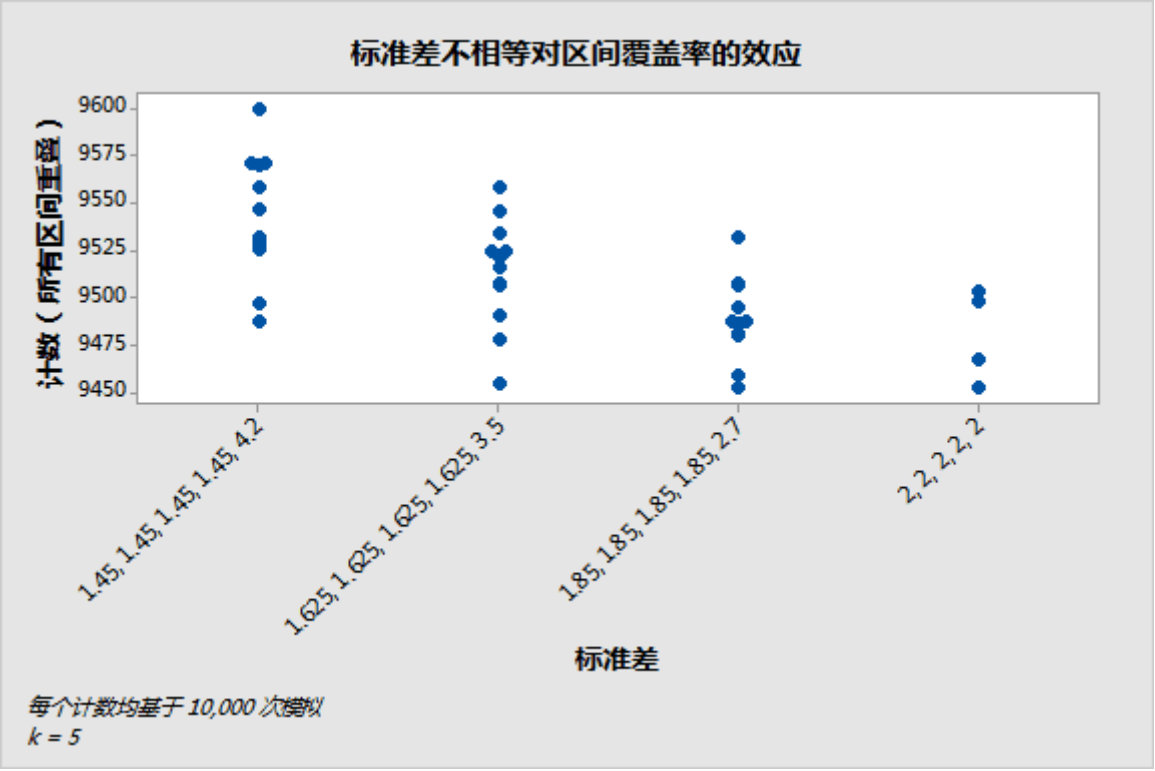


图 5　不同标准差的模拟结果

同时使用假设检验和比较区间

在极少数情况下，假设检验和比较可能不会同意否定原假设。假设检验可以否定原假设，比较区间则仍然全部重叠。相反，检验可能无法否定原假设，同时也存在不重叠的区间。这些分歧比较少见，因为这两种方法否定原假设成立的概率相同。

发生这种情况时，我们首先会考虑检验结果并使用比较在显著性检验事件中展开进一步的调查。如果检验在显著性水平 α 拒绝原假设，则无法与至少一个其他区间重叠的任何比较区间将被标记为红色。这可以用作直观提示，表明对应组的平均值至少与其他某个值不同。如果显著性检验指示“最有可能”差异，即使所有区间都重叠，重叠最少的平均值对也会被标记为红色（参见下面的图 6）。这个选择有点牵强，特别是当其他平均值对很少重叠时。但是其他平均值对并不限制其接近零的差值。

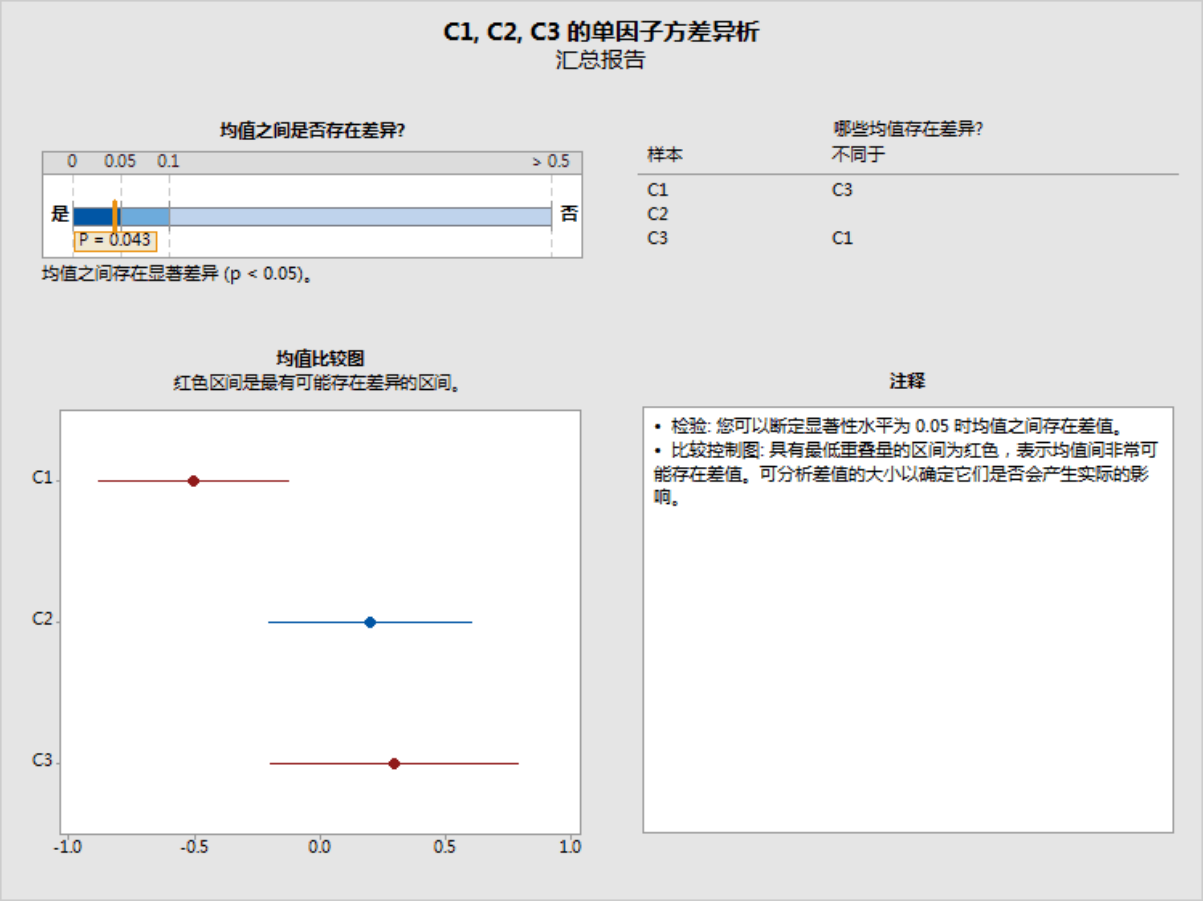


图 6　在显著性检验中，区间将被标记为红色，即使区间在样本中重叠也如此。

如果检验无法否定原假设，则任何区间都不会被标记为红色，即使存在不重叠的区间也如此（参见下面的图 7）。虽然区间表明平均值之间存在差异，但请记住，无法否定原假设并不能得出原假设成立的结论。它只表明，观测到的差值并没有大到足以排除偶然的原因。还要注意的是，在这种情况下，非重叠区间之间的差距通常非常小，因此非常小的差异仍与区间一致，而且这不一定表明存在具有实际意义的差异。

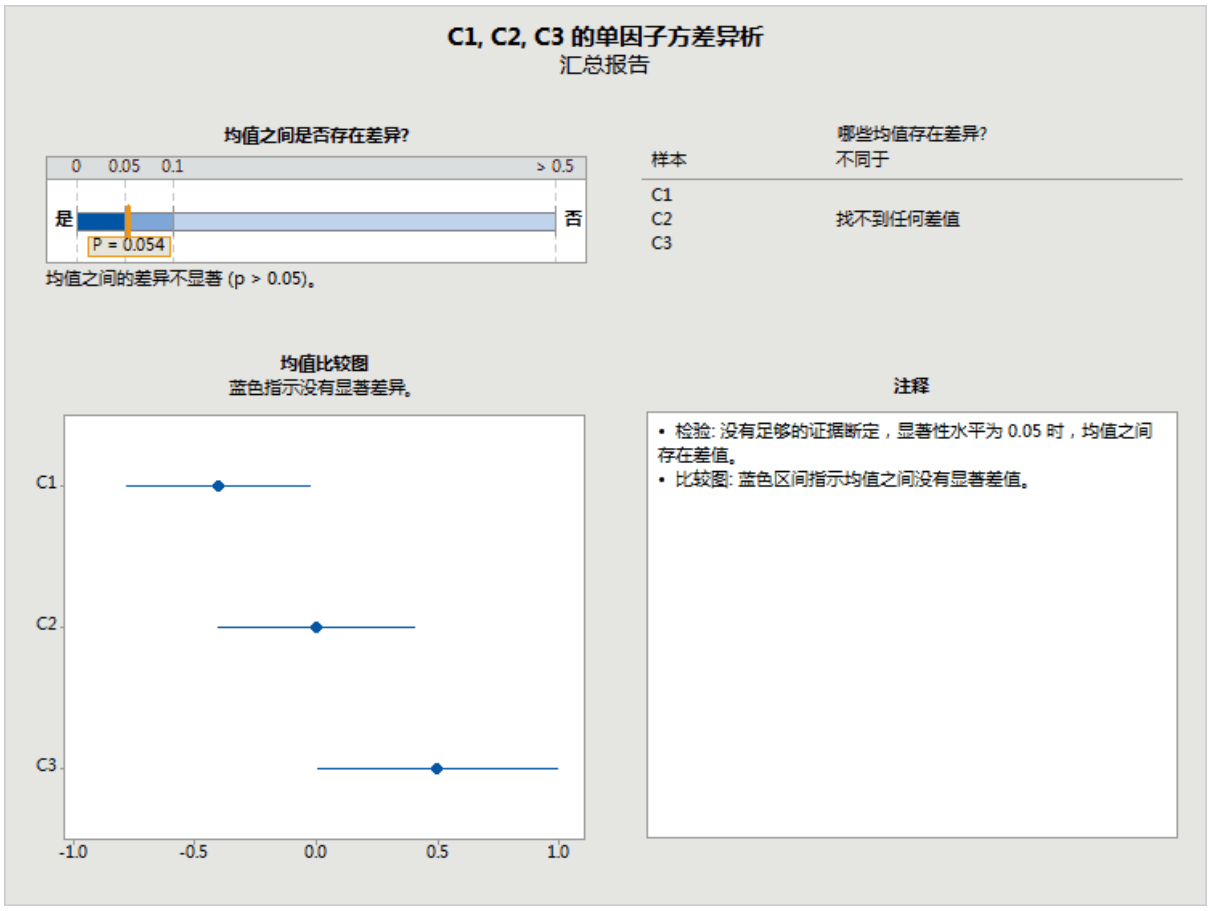


图 7　检验失败，没有一个区间被标记为红色，即使各个样本互不重叠也如此

附录 C：样本量

在单因子方差分析中，被检验的参数是不同组或总体的总体平均值 μ_1，μ_2，… μ_k。如果参数都相等，则满足原假设条件。如果平均值之间存在任何差异，则满足另一种假设条件。对于满足原假设条件的平均值，否定原假设的概率应不大于 α。实际概率取决于分布的标准差和样本量。检测到任何原假设偏差的功效会随标准差的减小或样本量的增大而增大。

我们可以使用非中心 F 分布计算出正态分布假设下具有相同标准差的 F 检验的功效。非中心参数为：

$$\theta_F = \sum_{i=1}^{k} n_i (\mu_i - \mu)^2 / \sigma^2$$

其中 μ 是加权平均值：

$$\mu = \sum_{i=1}^{k} n_i \mu_i / \sum_{i=1}^{k} n_i \ ,$$

σ 则是标准差，假定恒定不变。所有其他值都相等，功效随 θ_F 的增大而增大 确切地指，当平均值大幅偏离原假设时，功效将增大。

与 F 检验不同的是，Welch 检验没有一个既简单又准确的功效公式。不过，我们将来看一下两个相当不错的近似公式。第一个公式以类似于 F 检验的功效的方式使用非中心 F 分布。非中心参数仍将使用以下格式：

$$\theta_W = \sum_{i=1}^{k} w_i (\mu_i - \mu)^2$$

其中，μ 为加权平均值：

$$\mu = \sum_{i=1}^{k} w_i \mu_i / \sum_{j=1}^{k} w_j$$

但是加权值取决于标准差以及样本量，即 $w_i = n_i / \sigma_i^2$ 或者 $w_i = n_i / s_i^2$，这取决于我们是否模拟已知标准差 σ_i^2 的结果或者根据样本标准差 s_i^2 估计功效。然后根据以下方式计算近似功效：

$$P\left(F_{k-1,f,\theta_w} \geq F_{k-1,f,1-\alpha}\right)$$

其中分母自由度是

$$f = \frac{k^2 - 1}{3\sum_{i=1}^{k}\left(1 - w_i / \sum_{j=1}^{k} w_j\right)/(n_i - 1)}。$$

如下所示，这将为模拟中观测到的功效提供相当不错的近似值。而当我们在“协助”菜单中使用不同的近似值计算功效时，此菜单又能提供不错的观测窗口，从中可以选择平均值配置，我们将用这些平均值在“协助”菜单中计算功效。

平均值配置

为了与 Minitab（统计 > 方差分析 > 单因子）中使用的功效和样本量方法保持一致，“协助”不会要求用户提供用于评估功效的所有平均值。相反，它要求用户提供平均值之间一个具有实际意义的差值。对于给定的差值，可能有无限种平均值配置，最大平均值和最小平均值之间相差的数值也通过它们来决定。例如，下面一组 5 个平均值的最大差值全部为 10：

$\mu_1 = 0,\ \mu_2 = 5,\ \mu_3 = 5,\ \mu_4 = 5,\ \mu_5 = 10;$

$\mu_1 = 5,\ \mu_2 = 0,\ \mu_3 = 10,\ \mu_4 = 10,\ \mu_5 = 0;$

$\mu_1 = 0,\ \mu_2 = 10,\ \mu_3 = 0,\ \mu_4 = 0,\ \mu_5 = 0;$

这样的组合一定有很多。

我们采用在 Minitab （统计 > 功效和样本量 > 单因子方差分析）中用于检验功效和样本量的方法，在这种情况下，所有的平均值中只有两个平均值是（加权）平均值，其余两个平均值相差规定的数值。不过，由于方差和样本量可能不同，非中心参数（乃至功效）仍然取决于假定哪两个平均值不同。

考虑平均值 $\mu_1, \cdots, \mu_k$ 的配置，其中所有的平均值中只有两个平均值等于总加权平均值 μ，另外两个平均值（$\mu_i > \mu_j$）非但彼此不同，还不同于总平均值。用 $\Delta = \mu_i - \mu_j$ 表示两个平均值之间的差异。设置 $\Delta_i = \mu_i - \mu$ 且 $\Delta_j = \mu - \mu_j$。从而 $\Delta = \Delta_i + \Delta_j$。此外，由于 μ 代表所有 k 平均值的加权平均值，并且平均值的 (k - 2) 被认为等于 μ，我们得出：

$$\mu = \left[\sum_{l \neq i,j} w_l\mu_l + w_i(\mu + \Delta_i) + w_j(\mu - \Delta_j)\right] \Big/ \sum_{l=1}^{k} w_l = \mu + \left(w_i\Delta_i - w_j\Delta_j\right) \Big/ \sum_{l=1}^{k} w_l .$$

因此：

$$w_i\Delta_i = w_j\Delta_j = w_j(\Delta - \Delta_i),$$

因此，

$$\Delta_i = \frac{w_j}{w_i + w_j}\Delta$$

$$\Delta_j = \frac{w_i}{w_i + w_j}\Delta$$

对于这种特定的平均值配置，我们可以计算与 Welch 检验相关的非中心参数：

$$\theta_W = w_i(\mu_i - \mu)^2 + w_j(\mu_j - \mu)^2$$

$$= \frac{w_i w_j^2\Delta^2 + w_j w_i^2\Delta^2}{(w_i + w_j)^2} = \frac{w_i w_j\Delta^2}{w_i + w_j}$$

此值的 wj 如果固定不变，wi 会增大，反之亦然。因此具有两个最大的加权值的平均值对 (i, j) 最大，具有两个最小的加权值的平均值对 (i, j) 则最小。所有功效计算都会考虑这两种极端情况，假设恰好有两个平均值与总加权平均值不同，这将得到功效的最大值和最小值。

如果为检验指定差值，则将针对此差值评估功效的最小值和最大值。这些功效的范围显示在带彩色条的报告上，其中等于或低于 60% 的功效用红色表示，等于或高于 90% 的功效用绿色表示，而 60% 至 90% 之间的功效用黄色表示。报告卡结果取决于功效的范围落在此彩色条纹的哪个区间内。如果整个范围呈红色，则任意一对平均值组的功效小于或等于 60%，并且报告卡上显示的红色图标指示功效不足问题。如果整个范围呈绿色，则任意组的功效至少为 90%，报告卡上的绿色图标表示此时功效充足。所有其他情况都被视为中间情况，通过报告卡上的黄色图标表示。

在不符合绿色条件的情况下，“协助”会计算由用户指定的差值和观测到的样本标准差导致出现绿色情况的样本量。估计功效取决于由加权值 $w_i = n_i/s_i^2$. 决定的样本量。如果假定所有样本都具有相同的样本量，那么两个最小的加权值将对应于具有最大样本标准差的两个组。如果指定的差值位于具有最大可变性的两个组之间，“协助”会发现提供至少 90% 的功效的样

本量。因此，为所有组使用至少 90% 的样本量将导致所有功效值至少为 90%，从而符合绿色条件。

如果用户没有指定功效计算的差值，则“协助”会发现最大差值，此时已计算功效的最大范围值将是 60%。这个值被标记在红条和黄条之间的边界处，对应于 60% 的功效。它还会发现最小差值，此时已计算功效的最小范围值将是 90%。这个值被标记在黄条和绿条之间的边界处，对应于 90% 的功效。

功效计算

功效使用近似值计算，参见 Kulinskaya et al. (2003)：

定义：

$\lambda = \sum_{i=1}^{k} w_i(\mu_i - \mu)^2$，

$A = \sum_{i=1}^{k} h_i$，

$B = \sum_{i=1}^{k} w_i(\mu_i - \mu)^2(1 - w_i/W)/(n_i - 1)$，

$D = \sum_{i=1}^{k} w_i^2(\mu_i - \mu)^4/(n_i - 1)$，

$E = \sum_{i=1}^{k} w_i^3(\mu_i - \mu)^6/(n_i - 1)^2$。

Welch 统计值的前三个分子累积值 $\sum_{i=1}^{k} w_i(\bar{x}_i - \hat{\mu})^2$ 可估计为：

$\kappa_1 = k - 1 + \lambda + 2A + 2B$，

$\kappa_2 = 2(k - 1 + 2\lambda + 7A + 14B + D)$，

$\kappa_3 = 8(k - 1 + 3\lambda + 15A + 45B + 6D + 2E)$。

用 $F_{k-1,f,1-\alpha}$ 表示 F(k - 1, f) 分布的 (1 - α) 分位数。提醒 $W^* \geq F_{k-1,f,1-\alpha}$ 为在 α 值 Welch 检验中否定原假设的标准。

设置

$q = (k - 1)\left[1 + \frac{2(k-2)A}{k^2 - 1}\right] F_{k-1,f,1-\alpha}$，

$b = \kappa_1 - 2\kappa_2^2/\kappa_3$，

$c = \kappa_3/(4\kappa_2)$ [注：c 的表达式显示在 Kulinskaya et al. (2003)中，不带括号。]

$v = 8\kappa_2^3/\kappa_3^2$。

然后，Welch 检验的近似估计功效为：

$$P(\chi_v^2 \geq \frac{q - b}{c})$$

其中 χ_v^2 是自由度为 v 的卡方随机变量。

下面是根据 10,000 次模拟，对 2 个近似方法的功效与各种示例的模拟功效进行比较的结果。

表 3　与模拟功效相比的两个近似方法的功效计算

示例	Alpha	模拟功效	非中心 F	Kulinskaya et al.
μ’s: 0, 0, 0, -0.1724, 0.8276	0.10	0.1372	0.135702	0.135795
σ’s: 2, 2, 2, 2, 4	0.05	0.0739	0.072563	0.069512
n’s: 12, 12, 12, 12, 10	0.01	0.0195	0.016587	0.012538
μ’s: 0, 0, 0, -0.3448, 1.6552	0.10	0.2498	0.251064	0.257455
σ’s: 2, 2, 2, 2, 4	0.05	0.1574	0.153128	0.156215
n’s: 12, 12, 12, 12, 10	0.01	0.0541	0.045211	0.042195
μ’s: 0, 0, 0, -0.5172, 2.4828	0.10	0.4534	0.44557	0.453506
σ’s: 2, 2, 2, 2, 4	0.05	0.3211	0.311994	0.321575
n’s: 12, 12, 12, 12, 10	0.01	0.1273	0.121225	0.125065
μ’s: 0, 0, 0, -0.6896, 3.3104	0.10	0.662	0.671317	0.670296
σ’s: 2, 2, 2, 2, 4	0.05	0.5219	0.533819	0.538617
n’s: 12, 12, 12, 12, 10	0.01	0.2842	0.271316	0.282759
μ’s: 0, 0, 0, -0.8620, 4.1380	0.10	0.8417	0.852589	0.846697
σ’s: 2, 2, 2, 2, 4	0.05	0.7382	0.752173	0.746121
n’s: 12, 12, 12, 12, 10	0.01	0.4883	0.487601	0.49323
μ’s: 0, 0, 0, -1.0344, 4.9656	0.10	0.9429	0.952077	0.954929
σ’s: 2, 2, 2, 2, 4	0.05	0.8866	0.901485	0.897937
n’s: 12, 12, 12, 12, 10	0.01	0.691	0.711055	0.703379
μ’s: 0, 0, 0, 0, 0, -0.148148, 1.85185	0.10	0.2011	0.189392	0.200114
σ’s: 2, 2, 2, 2, 2, 2, 5	0.05	0.1201	0.108986	0.11742
n’s: 20, 20, 20, 20, 20, 20, 10	0.01	0.0385	0.028986	0.031456
μ’s: 0, 0, 0, 0, 0, -0.296296, 3.70370	0.10	0.4942	0.485917	0.500143
σ’s: 2, 2, 2, 2, 2, 2, 5	0.05	0.3677	0.351593	0.375296
n’s: 20, 20, 20, 20, 20, 20, 10	0.01	0.177	0.149041	0.177189
μ’s: 0, 0, 0, 0, 0, -0.444444, 5.55556	0.10	0.8125	0.829702	0.819542
σ’s: 2, 2, 2, 2, 2, 2, 5	0.05	0.7131	0.727384	0.720807
n’s: 20, 20, 20, 20, 20, 20, 10	0.01	0.4876	0.474291	0.49469
μ’s: 0, 0, 0, 0, 0, -0.592593, 7.40741	0.10	0.9645	0.977211	0.984213
σ’s: 2, 2, 2, 2, 2, 2, 5	0.05	0.9286	0.949997	0.949239
n’s: 20, 20, 20, 20, 20, 20, 10	0.01	0.7938	0.831174	0.814067

示例	Alpha	模拟功效	非中心 F	Kulinskaya et al.
μ’s: 0, 0, 0, 0, 0, -0.740741, 9.25926	0.10	0.9961	0.998947	1.00
σ’s: 2, 2, 2, 2, 2, 2, 5	0.05	0.9895	0.996653	1.00
n’s: 20, 20, 20, 20, 20, 20, 10	0.01	0.9528	0.977536	0.98705
μ’s: 0, 0, 0, 0, 0, -0.888889, 11.1111	0.10	0.9999	0.999985	1.00
σ’s: 2, 2, 2, 2, 2, 2, 5	0.05	0.9995	0.999926	1.00
n’s: 20, 20, 20, 20, 20, 20, 10	0.01	0.9943	0.99891	1.00
μ’s: 0, 0, 0, 0, 0, -0.518519, 6.48148	0.10	0.9059	0.929392	0.924696
σ’s: 2, 2, 2, 2, 2, 2, 5	0.05	0.8403	0.868721	0.85672
n’s: 20, 20, 20, 20, 20, 20, 10	0.01	0.6511	0.67121	0.66652
μ’s: 0, 0, 0, 0, 0, -.5, .5	0.10	0.187	0.186658	0.18329
σ’s: 2, 2, 2, 2, 2, 2, 2	0.05	0.1098	0.106600	0.100189
n’s: 12, 12, 12, 12, 12, 12, 12	0.01	0.0315	0.027773	0.021332
μ’s: 0, 0, 0, 0, 0, -1, 1	0.10	0.4734	0.474736	0.472469
σ’s: 2, 2, 2, 2, 2, 2, 2	0.05	0.3394	0.338655	0.33443
n’s: 12, 12, 12, 12, 12, 12, 12	0.01	0.1378	0.137788	0.128693
μ’s: 0, 0, 0, 0, 0, -1.5, 1.5	0.10	0.8228	0.817355	0.810181
σ’s: 2, 2, 2, 2, 2, 2, 2	0.05	0.7112	0.707319	0.698461
n’s: 12, 12, 12, 12, 12, 12, 12	0.01	0.4391	0.441154	0.431868
μ’s: 0, 0, 0, 0, 0, -2, 2	0.10	0.9691	0.973246	0.973319
σ’s: 2, 2, 2, 2, 2, 2, 2	0.05	0.9312	0.940585	0.936546
n’s: 12, 12, 12, 12, 12, 12, 12	0.01	0.7817	0.799339	0.785099
μ’s: 0, 0, 0, 0, 0, -2.5, 2.5	0.10	0.9984	0.998579	0.999763
σ’s: 2, 2, 2, 2, 2, 2, 2	0.05	0.9936	0.99533	0.997481
n’s: 12, 12, 12, 12, 12, 12, 12	0.01	0.9587	0.967674	0.966249
μ’s: 0, 0, 0, 0, 0, -3, 3	0.10	1.00	0.999975	1.00
σ’s: 2, 2, 2, 2, 2, 2, 2	0.05	0.9997	0.99987	1.00
n’s: 12, 12, 12, 12, 12, 12, 12	0.01	0.9959	0.997927	0.99961
μ’s: 0, 0, 0, 0, 0, -3.5, 3.5	0.10	1.00	1.00	1.00
σ’s: 2, 2, 2, 2, 2, 2, 2	0.05	1.00	1.00	1.00
n’s: 12, 12, 12, 12, 12, 12, 12	0.01	0.99998	0.99995	1.00
μ’s: 0, 0, 0, 0, 0, -1.75, 1.75	0.10	0.914	0.921225	0.916652
σ’s: 2, 2, 2, 2, 2, 2, 2	0.05	0.8418	0.852755	0.843856
n’s: 12, 12, 12, 12, 12, 12, 12	0.01	0.619	0.633815	0.620704

示例	Alpha	模拟功效	非中心 F	Kulinskaya et al.
μ’s: 0, -0.5, 0.5	0.10	0.2548	0.259249	0.257149
σ’s: 2, 2, 2	0.05	0.1549	0.160861	0.156251
n’s: 12, 12, 12	0.01	0.0470	0.049045	0.042292
μ’s: 0, -1, 1	0.10	0.654	0.659073	0.654105
σ’s: 2, 2, 2	0.05	0.5205	0.522885	0.515816
n’s: 12, 12, 12	0.01	0.2612	0.26355	0.252469
μ’s: 0, -1.5, 1.5	0.10	0.9364	0.935939	0.937768
σ’s: 2, 2, 2	0.05	0.8747	0.87562	0.872608
n’s: 12, 12, 12	0.01	0.6614	0.664478	0.652563
μ’s: 0, -1.75, 1.75	0.10	0.981	0.981434	0.986815
σ’s: 2, 2, 2	0.05	0.9522	0.9561	0.959796
n’s: 12, 12, 12	0.01	0.8251	0.830726	0.823624
μ’s: 0, -2, 2	0.10	0.9953	0.995969	0.999332
σ’s: 2, 2, 2	0.05	0.9878	0.988175	0.993705
n’s: 12, 12, 12	0.01	0.9308	0.931922	0.933446
μ’s: 0, -2.5, 2.5	0.10	0.9999	0.999923	1.00
σ’s: 2, 2, 2	0.05	0.9997	0.999634	1.00
n’s: 12, 12, 12	0.01	0.9949	0.994725	0.99909
μ’s: 0, -3, 3	0.10	1.00	1.00	1.00
σ’s: 2, 2, 2	0.05	1.00	1.00	1.00
n’s: 12, 12, 12	0.01	0.9999	0.99985	1.00
μ’s: 0, -3.5, 3.5	0.10	1.00	1.00	1.00
σ’s: 2, 2, 2	0.05	1.00	1.00	1.00
n’s: 12, 12, 12	0.01	0.9999	1.00	1.00
μ’s: 0, -0.142857, 0.857143	0.10	0.1452	0.143156	0.146824
σ’s: 2, 2, 4	0.05	0.0790	0.077699	0.077538
n’s: 14, 12, 8	0.01	0.0223	0.018200	0.014338
μ’s: 0, -0.285714, 1.71429	0.10	0.2765	0.27424	0.286222
σ’s: 2, 2, 4	0.05	0.1787	0.170628	0.179469
n’s: 14, 12, 8	0.01	0.0624	0.051588	0.050335
μ’s: 0, -0.428571, 2.57143	0.10	0.4861	0.476925	0.490018
σ’s: 2, 2, 4	0.05	0.3487	0.338626	0.355743
n’s: 14, 12, 8	0.01	0.1467	0.132405	0.141352

示例	Alpha	模拟功效	非中心 F	Kulinskaya et al.
μ’s: 0, -0.50000, 3	0.10	0.5846	0.588533	0.596795
σ’s: 2, 2, 4	0.05	0.4425	0.444491	0.460707
n’s: 14, 12, 8	0.01	0.2107	0.19729	0.212798
μ’s: 0, -0.571429, 3.42857	0.10	0.6933	0.694684	0.696773
σ’s: 2, 2, 4	0.05	0.5631	0.555731	0.567129
n’s: 14, 12, 8	0.01	0.3052	0.279131	0.299302
μ’s: 0, -0.714286, 4.28571	0.10	0.848	0.861469	0.859329
σ’s: 2, 2, 4	0.05	0.7402	0.759703	0.759762
n’s: 14, 12, 8	0.01	0.4871	0.480052	0.497421
μ’s: 0, -0.857143, 5.14286	0.10	0.9434	0.952562	0.961913
σ’s: 2, 2, 4	0.05	0.8869	0.898817	0.902716
n’s: 14, 12, 8	0.01	0.6649	0.687058	0.692591
μ’s: 0, -1, 6	0.10	0.9849	0.987981	0.999989
σ’s: 2, 2, 4	0.05	0.9609	0.967589	0.985049
n’s: 14, 12, 8	0.01	0.8294	0.847436	0.853787
μ’s: 0, -1.14286, 6.85714	0.10	0.9976	0.997776	1.00
σ’s: 2, 2, 4	0.05	0.989	0.99222	1.00
n’s: 14, 12, 8	0.01	0.9222	0.940972	0.96383
μ’s: 1, 2, 3	0.10	0.8838	0.882194	0.884649
σ’s: 0.3, 2.4, 3.6	0.05	0.7995	0.797869	0.802137
n’s: 13, 19, 25	0.01	0.5632	0.556486	0.563208
μ’s: 1, 2, 3	0.10	0.5649	0.566831	0.565141
σ’s: 2.77489, 2.77489, 2.77489	0.05	0.4305	0.431302	0.428126
n’s: 13, 19, 25	0.01	0.1994	0.201329	0.195734

上述结果在下图中总结，其中显示了各个近似值与通过模拟估计的功效值之间的差异。

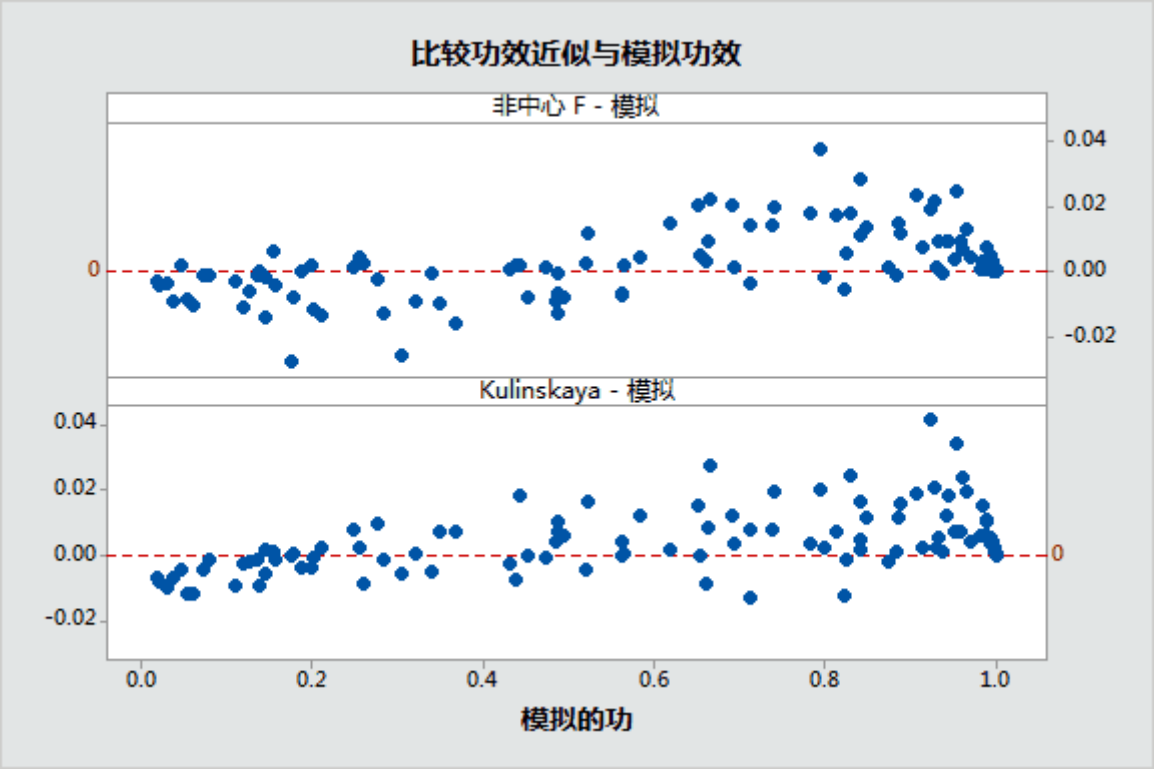


图 8　两个功效近似值与通过模拟估计的功效比较

附录 D：正态性

在本节中，我们提出了检查多个非正态分布中的小到中等样本的 Welch 检验和比较区间的性能模拟。

下表总结了在平均值相等的原假设下，不同类型的分布的模拟结果。对于这些示例，所有标准差也都相等，而且所有样本量都相等。样本数 k = 3、5 或 7。

每个单元格显示基于 10,000 次模拟的 I 类错误的估计值。目标显著性水平（目标值 α）为 0.05。

表 4　不同分布平均值相等的 Welch 检验的模拟结果

	样本量 n = 10			**样本量 n = 15**		
分布	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7
N(0,1)	0.0490	0.0486	0.0512	0.0534	0.0522	0.0550
T(3)	0.0371	0.0361	0.0348	0.0353	0.0385	0.0365
T(5)	0.0440	0.0425	0.0439	0.0435	0.0428	0.0428
Laplace(0,1)	0.0433	0.0354	0.0345	0.0445	0.0397	0.0407
Uniform(-1, 1)	0.0544	0.0640	0.0718	0.0517	0.0573	0.0585
Beta(3, 3)	0.0504	0.0577	0.0622	0.0501	0.0538	0.0564
指数	0.0508	0.0621	0.0748	0.0483	0.0633	0.0779
卡方(3)	0.0473	0.0579	0.0753	0.0499	0.0588	0.0703
卡方(5)	0.0458	0.0594	0.0643	0.0504	0.0606	0.0679
卡方(10)	0.0463	0.0510	0.0585	0.0463	0.0552	0.0567
Beta(8, 1)	0.0500	0.0622	0.0775	0.0549	0.0653	0.0760

I 类错误概率都在目标值 α 的 3 个百分点范围内，即使样本量为 10 也如此。较大的偏差往往会出现在更多的组和非正态分布中。对于 10 个样本，仅当 k =7 时，接受概率会相差 2 个百分点以上。这些发生在尾部比正态分布更短的均匀分布，以及高度偏态指数、卡方 (3) 和 beta(8, 1) 分布中。将样本量增加到 15 个可显著改进均匀分布的结果，但这不适用于两个重度偏态分布。

我们对比较区间进行了一项类似的模拟。在这种情况下，模拟 α 超过了 10,000 次，其中的一些区间并不重叠。目标值 $\alpha = 0.05$。

表 5　不同分布平均值相等的比较区间的模拟结果

	样本量 n = 10			样本量 n = 15		
分布	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7
N(0,1)	0.0493	0.0494	0.0469	0.0538	0.0518	0.0561
t(3)	0.0378	0.0321	0.0254	0.0347	0.0343	0.0289
t(5)	0.0449	0.0399	0.0361	0.0447	0.0444	0.0412
Laplace(0,1)	0.0438	0.0305	0.0246	0.0456	0.0366	0.0348
Uniform(-1, 1)	0.0559	0.0605	0.0699	0.0534	0.0607	0.0590
Beta(3, 3)	0.0515	0.0569	0.0615	0.0510	0.0553	0.0568
指数	0.0353	0.0254	0.0207	0.0346	0.0310	0.0275
卡方(3)	0.0375	0.0305	0.0296	0.0384	0.0359	0.0339
卡方(5)	0.0405	0.0390	0.0353	0.0417	0.0433	0.0416
Chi-square(10)	0.0425	0.0428	0.0447	0.0435	0.0476	0.0464
Beta(8, 1)	0.0381	0.0352	0.0287	0.0459	0.0428	0.0403

在 Welch 检验中，I 类错误概率都在目标值 α 的 3 个百分点范围内，即使样本量为 10 时也如此。较大的偏差往往会出现在更多的样本和非正态分布中。对于 10 个样本，当 k = 7 时，错误概率有时会相差 2 个百分点以上（当 k= 5 时也如此）。这些情况会在自由度为 3 的极值重尾 t 分布、Laplace 分布，以及高度偏态指数与卡方 (3) 分布中发生。将样本量增加到 15 个将改进上述结果，从而只有 t(3) 和指数分布的模拟值 α 偏离目标值 2 个百分点以上。请注意，与 Welch 检验结果不同的是，较大的比较区间偏差将趋于保守。

“协助”中的单因子方差分析允许最多 k=12 的样本，因此接下来我们将考虑超过 7 个样本的结果。下表将显示对 K=9 组中的非正态数据使用 Welch 检验得到的 I 类错误概率。这一次的目标值又是 $\alpha = 0.05$。

表 6　对 9 个样本的不同分布执行 Welch 检验的模拟结果

分布	k = 9
t(3)	0.0362
t(5)	0.0426
Laplace(0,1)	0.0402
Uniform(-1, 1)	0.0625
Beta(3, 3)	0.0584

指数	0.0885
卡方(3)	0.0774
卡方(5)	0.0686
Chi-square(10)	0.0581
Beta(8, 1)	0.0863

正如预料的那样，重度偏态分布表明目标值 α 的最大偏差。即便如此，任何一个错误概率都不会偏离目标值 4 个百分点以上，虽然指数分布的偏差是如此接近。报告卡可处理 15 个样本，这将无法指出非正态数据问题，因为所有的结果至少都相当接近目标值 α。

样本量 n = 15 时不如样本量 k = 12 时有效。下面，我们将考虑对一系列使用极值非正态分布的样本量执行 Welch 检验的模拟结果，这将帮助我们制定一个合理的样本量标准。

表 7　对 12 个样本的不同分布执行 Welch 检验的模拟结果

n	T(3)	均匀	卡方(5)
10	0.0397	0.0918	0.0792
15	0.0351	0.0695	0.0717
20	0.0362	0.0622	0.0671
30	0.0408	0.0573	0.0657

对于这些分布，如果我们愿意接受略微偏离目标值 α 2 个百分点以上，则表示接受 n = 15。为了确保偏差低于 2 个百分点，样本量应为 20。现在，我们将考虑从重度偏态卡方 (3) 和指数分布中得到的结果。

表 8　对 12 个样本的卡方分布和指数分布执行 Welch 检验的模拟结果

n	卡方(3)	指数
10	0.1013	0.1064
15	0.0854	0.1079
20	0.0850	0.0951
30	0.0746	0.0829
40	0.0727	0.0735
50	0.0675	0.0694

这些重度偏态分布将带来更大的挑战。如果我们要接受正好偏离目标值 $\alpha = 0.05$ 3 个百分点的偏差，则可以考虑 n = 15，即使是卡方 (3) 分布也能满足此条件，不过指数分布却要求使用更接近的 n = 30。当特定的样本量标准有点牵强，而且 n=20 十分贴合范围广泛的分布要

求，同时略微符合极值偏态分布要求时，我们使用 n = 20 作为 10～12 个样本的最小推荐样本量。显然，如果需要保持较小的偏差（即使对于极值偏态分布也是如此），则建议使用较大的样本。