双样本 t 检验

概述

可使用双样本 t 检验来比较两个独立组是否存在差异。此检验是在总体呈正态分布并且具有相等方差的假设情况下推导出的。虽然正态性假设并不是至关重要（Pearson，1931；Barlett，1935；Geary，1947），但如果样本数量差异明显，则等方差假设至关重要（Welch，1937；Horsnell，1953）。

有些工作人员首先执行初步检验以评估等方差，然后才执行经典的双样本 t 过程。但此方法具有严重缺点，因为这些方差检验受重要的假设和限制约束。例如，许多等方差检验（如经典 F 检验）对正态性偏离敏感。不依赖于正态性假设的其他检验（如 Levene/Brown-Forsythe）在检测方差间差值方面的功效比较低。

B.L. Welch 开发了在方差不必相等时，用于比较两个独立正态总体的均值的近似方法（Welch，1947）。由于 Welch 的修正的 t 检验不是在等方差假设下推导出的，因此，它可让用户比较两个总体的均值，而不必先检验是否存在等方差。

在本白皮书中，我们将 Welch 修改的 t 方法与经典的双样本 t 过程进行比较，并确定哪个过程最可靠。我们还介绍了在“协助”的“报告卡”中自动执行和显示的以下数据检查，并说明它们对分析结果的影响：

- 正态性
- 异常数据
- 样本数量

双样本 t 检验方法

经典的双样本 t 检验与 Welch 的 t 检验

如果数据来自具有相同方差的两个正态总体，则经典的双样本 t 检验的功效高于或等于 Welch 的 t 检验。正态性假设对于经典过程（Pearson，1931；Barlett，1935；Geary，1947）并不是至关重要，但等方差对于确保有效结果却很重要。具体而言，当样本数量不同时，不论样本数量多大，经典过程都对等方差假设敏感（Welch，1937；Horsnell，1953）。但实际上，等方差假设很少为真，这会导致更高的 I 类错误率。因此，如果在这两个样本具有不同的方差时使用经典的双样本 t 检验，则此检验很可能产生不正确的结果。

Welch 的 t 检验是经典的 t 检验的可行替代方法，因为它不假设等方差，因此对方差不等的不敏感是对所有样本数量都成立的。但是，Welch 的 t 检验基于近似方法，小样本的性能可能会有问题。我们想要确定 Welch 的 t 检验和经典的双样本 t 检验中的哪一个是“协助”中最可靠、最实用的检验。

目标

我们想要通过模拟研究和理论推导确定是 Welch 的 t 检验还是经典的双样本 t 检验更可靠。具体地说，我们想要检查：

- 当数据呈正态分布并且方差相等时，对不同的样本数量进行经典的双样本 t 检验和 Welch 的 t 检验的 I 类和 II 类错误率。
- 针对经典的双样本 t 检验失效的不平衡和不等方差设计的 Welch 的 t 检验的 I 类和 II 类错误率。

方法

我们的模拟围绕以下三个方面展开：

- 我们在不同模型假设（包括正态性、非正态性、等方差、不等方差、平衡和不平衡设计）下比较了经典的双样本 t 检验和 Welch 的 t 检验的模拟检验结果。有关详细信息，请参见附录 A。
- 我们推导出 Welch 的 t 检验的功效函数，并将其与经典的双样本 t 检验的功效函数进行了比较。有关详细信息，请参见附录 B。
- 我们研究了非正态性对 Welch 的 t 检验的理论功效函数的影响。

结果

在经典的双样本 t 模型假设为真时，Welch 的 t 检验与经典的双样本 t 检验的效果一样或接近，但小样本的不平衡设计除外。但是，当设计为小样本且不平衡时，因为对等方差假设敏感，经典的双样本 t 检验的效果也不佳。此外，在实际环境中，很难确定两个总体具有完全一样的方差。因此，经典的双样本检验的理论优势与 Welch 的 t 检验几乎持平。鉴于此原因，“协助”使用 Welch 的 t 检验来比较两个总体的均值。有关详细的模拟结果，请参见附录 A、B 和 C。

数据检查

正态性

Welch 的 t 检验是在“协助”中用于比较两个独立总体的均值的方法，是在总体呈正态分布的假设下推导出的。幸运的是，即使数据不呈正态分布，只要样本足够大，Welch 的 t 检验效果就不错。

目标

我们想要确定 Welch 方法和经典的双样本 t 检验的模拟显著性水平与目标显著性水平（I 类错误率）0.05 的匹配程度。

方法

我们对源自正态、偏斜和污染正态（等方差和不等方差）总体的 10,000 对独立样本执行了 Welch 的 t 检验和经典的双样本 t 检验模拟。样本数量各不相同。正态总体可作为对照总体使用，以便进行比较。对于每种条件，我们计算了模拟显著性水平，并将其与 0.05 的目标或名义显著性水平进行了比较。如果此检验执行效果不错，则模拟显著性水平应接近 0.05。

结果

对于中等或大样本，Welch 的 t 检验可维持正态和非正态数据的 I 类错误率。如果这两个样本数量至少为 15，则模拟显著性水平接近于目标显著性水平。有关详细信息，请参见附录 A。

由于采用相对较小的样本检验执行效果不错，因此，“协助”不会检验数据的正态性， 而是检查样本数量，并在“报告卡”中显示以下状态指示符：

状态	条件
	这两个样本数量至少为 15；正态性不是问题。
	至少有一个样本数量小于 15；正态性可能是个问题。

异常数据

异常数据是极大或极小的数据值，也称作离群值。异常数据可能会对分析结果造成较强的影响。在样本较小时，可能会影响找到统计意义显著结果的机会。异常数据可表示数据收集出现问题，或过程的异常行为。因此，这些数据点通常值得调查，并在可能时给予校正。

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离群值的方法开发了用于检查离群值的方法。

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离群值。下四分位和上四分位是数据的第 25 个和第 75 个四分位数。四分位间距是两个四分位之间的差值。即使存在多个离群值，此方法的效果也不错，因为它可以检测到每个特定的离群值。

仅当样本数量非常小时，离群值往往会对功效函数产生影响。通常，在出现离群值时，观测到的功效值往往略大于目标理论功效值。可以在附录 C 的图 10 中看到这种模式，其中的模拟和理论功效曲线并不十分接近，直到最小样本数量达到 15。

在检查异常数据时，“协助”的双样本 t 检验“报告卡”中会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，可能会影响检验结果。

样本数量

通常，为收集拒绝“无差异”原假设的证据，将进行假设检验。如果样本太小，检验的功效可能不准确，因此检测不到均值之间实际存在的差值，这将导致 II 类错误。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

如果当前数据没有提供足够的证据拒绝原假设，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有实际意义的差值。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，但样本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

功效和样本数量分析基于用于执行统计分析的特定检验的理论功效函数。对于 Welch 的 t 检验，此功效函数取决于样本数量、两个总体均值之间的差值和两个总体的真方差。有关详细信息，请参见附录 B。

结果

在数据不能提供足够的证据拒绝原假设时，“协助”将计算出在给定样本数量的条件下，能以 80% 和 90% 的概率检测到的实际差值。此外，如果用户提供了相关的实际差值，则“协助”将计算有 80% 和 90% 的机会检测到此实际差值的样本数量。

由于结果取决于用户的特定样本，因此，没有要报告的一般结果。但是，有关 Welch 的检验功效函数的详细信息，请参考附录 B 和 C。

“协助”的双样本 t 检验“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
	此检验发现均值之间存在差值，因此，功效不是问题。 或 功效足够。此检验没有发现均值之间存在差值，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。此检验没有发现均值之间存在差值，但样本数量足够大，有 80% 到 90% 的机会检测到给定差值。将报告达到 90% 功效所需的样本数量。
	功效可能不足。此检验没有发现均值之间存在差值，但样本数量足够大，有 60% 到 80% 的机会检测到给定差值。将报告达到 80% 和 90% 功效所需的样本数量。
	功效不足。此检验没有发现均值之间存在差值，并且样本数量不够大，无法提供至少 60% 的机会检测到给定差值。将报告达到 80% 和 90% 功效所需的样本数量。
	此检验没有发现均值之间存在差值。您没有指定要检测的均值之间的实际差值，因此，此报告根据您的样本数量、标准差和 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.

Aspin, A. A. (1949). Tables for Use in Comparisons whose Accuracy Involves Two Variances, Separately Estimated, Biometrika, 36, 290-296.

Bartlett, M. S. (1935). The effect of non-normality on the t-distribution. *Proceedings of the Cambridge Philosophical Society, 31, 223-231.*

Box, G. E. P. (1953). Non-normality and Tests on Variances, Biometrika, 40, 318-335.

Geary, R. C. (1947). Testing for Normality, Biometrika, 34, 209-242.

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.*

Horsnell, G. (1953). The effect of unequal group variances on the F test for homogeneity of group means. *Biometrika, 40, 128-136.*

James, G. S. (1951). The comparison of several groups of observations when the ratios of the populations variances are unknown, Biometrika, 38, 324-329.

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.

Lehmann, E. L. (1959). *Testing statistical hypotheses*. New York, NY: Wiley.

Neyman, J., Iwaszkiewicz, K. & Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation, Journal of the Royal Statistical Society, Series B, 2, 107-180.

Pearson, E. S. (1931). The Analysis of variance in case of non-normal variation, Biometrika, 23, 114-133.

Pearson, E.S. & Hartley, H.O. (Eds.). (1954). *Biometrika Tables for Statisticians*, Vol. I. London: Cambridge University Press.

Srivastava, A. B. L. Effect of non-normality on the power function of t-test, Biometrika, 45, 421-429.

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

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

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal, Biometrika, 29, 350-362.

Wolfram, S. (1999). *The Mathematica Book (*4th ed.). Champaign, IL: Wolfram Media/Cambridge University Press.

附录 A：非正态性和非齐性对经典的双样本 t 检验和 Welch t 检验的影响

我们进行了几项模拟研究，这些研究专为在不同模型假设情况下比较经典的双样本 t 检验和 Welch 的 t 检验而设计。

模拟研究 A

我们分三个部分进行了研究：

- 在此研究的第一部分中，我们分析了在正态性假设为真时，经典的双样本 t 检验和 Welch 的 t 检验对等方差假设的敏感度。从两个独立的正态总体中生成了两个样本。第一个样本为基础样本，是从均值为 0 且标准差为 $\sigma_1 = 2$ 的正态总体中抽取的 $N(0,2)$。第二个样本也是从均值为 0 的正态总体中抽取的，但选择了标准差 σ_2，从而比率 $\rho = \sigma_2/\sigma_1$ 为 0.5、1.0、1.5 和 2。换言之，第二个样本分别从总体 $N(0,1)$、$N(0,2)$、$N(0,3)$ 和 $N(0,4)$ 中抽取得到。此外，每种情况下的基本样本数量固定为 $n_1 = 5,10,15,20$，对于每个给定的 n_1，选定了第二个样本数量 n_2，这样，样本数量比率 $r = n_2/n_1$ 约等于 0.5、1、1.5 和 2.0。

  对于这些双样本设计中的每个设计，我们根据各个总体生成了 10,000 对独立样本。然后，我们对这 10,000 对样本中的每对样本执行了经典的双样本 t 检验和 Welch 的 t 检验，以检验均值之间无差值的原假设。由于均值之间的真实差值为零，否定原假设的 10,000 个仿行中的部分仿行表示检验的模拟显著性水平。由于每个检验的目标显著性水平为 $\alpha = 0.05$，因此，与每个检验和每个试验关联的模拟误差大约为 0.2%。
- 在第二部分中，我们调查了非正态性（尤其是偏斜度）对这两个检验的模拟显著性水平的影响。此模拟的设置方法与上一模拟的设置方法相同，但不同之处是，基本样本是从自由度为 2（即 Chi(2)）的卡方分布中抽取的，第二个样本是从其他卡方分布中抽取的，这样，$\rho = \sigma_2/\sigma_1$ 采用值 0.5、1.0、1.5 和 2。均值之间的假设差值设置为父总体均值之间的真实差值。
- 在第三部分中，我们检查了离群值对这两个 t 检验的性能的影响。鉴于此原因，这两个样本都是从污染正态分布中抽取的。污染正态总体 $\mathrm{CN}(\mathrm{p},\sigma)$ 是下面两个正态总体的混合：$\mathrm{N}(0,1)$ 总体和正态 $\mathrm{N}(0,\sigma)$ 总体。我们将污染正态分布定义为：

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

其中，p 是混合参数，$1-\mathrm{p}$ 是污染比例或离群值比例。很容易发现，如果 X 作为 $CN(p,\sigma)$ 分布，则其均值为 $\mu_X = 0$，其标准差为 $\sigma_X = \sqrt{p + (1-p)\sigma^2}$。

基本样本是从 $CN(.8,4)$ 中抽取的，第二个样本是从污染正态总体 $CN(.8,\sigma)$ 中抽取的。选择了参数 σ，以使两个（污染）总体的标准差比值 $\rho = \sigma_2/\sigma_1$ 等于 0.5、1.0、1.5 和 2，与第 1 部分和第 II 部分类似。由于 $\sigma_1 = \sqrt{.8 + (1-.8) * 16} = 2.0$，这将导致分别选择 $\sigma = 1,4,6.40,8.72$。换言之，第二个样本是从 $CN(.8,1)$、$CN(.8,4)$、$CN(.8,6.4)$ 和 $CN(.8,8.72)$ 中抽取的。然后，我们执行了第 I 部分中所述的模拟。

研究结果已在表 1 中进行了整理，并显示在图 1、图 2 和图 3 中。

结果与汇总

通常，在正态性和等方差假设条件下，模拟结果支持理论结果，经典的双样本 t 检验会产生与目标水平接近的显著性水平，即使在样本数量很小时也是如此。图 1 中控制图的第二列显示两个正态总体方差相等的设计中的模拟显著性水平。基于经典的双样本 t 检验的模拟显著性水平曲线无法与目标水平线区分开来。

下表显示了经典的双样本 t 检验和 Welch 的 t 检验（每个检验都具有 α = 0.05，基于根据正态总体、偏斜总体（卡方）和污染正态总体生成的样本对）的双侧检验的模拟显著性水平。这些样本对来自相同的分布族，但各个父总体的方差不一定相等。

表 1 n = 5 时，双侧检验（经典的双样本 t 检验和 Welch 的 t 检验，每个检验都具有 α = 0.05）的模拟显著性水平。

			基本总体：N(0,2) 第二个总体：N(0, σ_2)				基本总体：Chi(2) 第二个总体：卡方				基本总体：CN(.8,4) 第二个总体：CN(.8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$	方法	$n_1 = 5$				$n_1 = 5$				$n_1 = 5$			
3	.6	2T	.035	.050	.079	.105	.058	.042	.078	.113	.031	.036	.035	.034
		Welch	.035	.039	.049	.055	.048	.029	.055	.063	.029	.024	.021	.020
5	1.0	2T	.061	.052	.054	.058	.086	.036	.054	.064	.035	.031	.025	.023
		Welch	.048	.042	.044	.047	.066	.021	.040	.050	.027	.023	.018	.016
8	1.6	2T	.096	.048	.033	.027	.133	.041	.033	.032	.059	.037	.029	.024
		Welch	.050	.045	.043	.042	.094	.034	.032	.041	.034	.029	.026	.022
10	2.0	2T	.118	.055	.034	.025	.139	.041	.028	.024	.073	.041	.028	.023
		Welch	.052	.051	.050	.051	.097	.041	.033	.042	.035	.032	.028	.025

表 2　n = 10 时，双侧检验（经典的双样本 t 检验和 Welch 的 t 检验，每个检验都具有 α = 0.05 ）的模拟显著性水平

			基本总体：N(0,2) 第二个总体：N(0, σ_2)				基本总体：Chi(2) 第二个总体：卡方				基本总体：CN(.8,4) 第二个总体：CN(.8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$	方法	$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
5	.5	2T	.020	.050	.081	.112	.039	.044	.091	.123	.021	.035	.045	.047
		Welch	.046	.048	.050	.050	.043	.047	.067	.063	.034	.028	.022	.019
10	1.0	2T	.057	.051	.053	.055	.068	.044	.053	.054	.043	.042	.037	.032
		Welch	.051	.049	.049	.049	.062	.037	.046	.049	.039	.038	.032	.027
15	1.5	2T	.088	.048	.034	.029	.100	.043	.032	.032	.064	.040	.028	.021
		Welch	.050	.048	.047	.048	.074	.044	.041	.046	.035	.037	.035	.031
20	2	2T	.110	.048	.026	.019	.133	.042	.026	.022	.093	.046	.029	.019
		Welch	.048	.047	.045	.046	.083	.050	.044	.049	.036	.039	.040	.038

表 3　n = 15 时，双侧检验（经典的双样本 t 检验和 Welch 的 t 检验，每个检验都具有 α = 0.05）的模拟显著性水平

			基本总体：N(0,2) 第二个总体：N(0, σ_2)				基本总体：Chi(2) 第二个总体：卡方				基本总体：CN(.8,4) 第二个总体：CN(.8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$	方法	$n_1 = 15$				$n_1 = 15$				$n_1 = 15$			
8	.53	2T	.021	.050	.083	.110	.036	.041	.089	.114	.022	.044	.056	.062
		Welch	.050	.051	.051	.050	.047	.049	.067	.062	.044	.036	.027	.022
15	1.0	2T	.049	.047	.050	.053	.064	.046	.051	.061	.045	.045	.041	.037
		Welch	.045	.046	.049	.048	.060	.042	.048	.057	.042	.043	.039	.033
23	1.53	2T	.081	.049	.033	.028	.103	.042	.036	.030	.075	.048	.033	.024
		Welch	.048	.049	.048	.050	.071	.042	.048	.050	.042	.045	.044	.041
30	2.0	2T	.111	.050	.028	.018	.123	.049	.027	.020	.100	.046	.025	.016

			基本总体：N(0,2) 第二个总体：N(0, σ_2)				基本总体：Chi(2) 第二个总体：卡方				基本总体：CN(.8,4) 第二个总体：CN(.8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$	方法	$n_1 = 15$				$n_1 = 15$				$n_1 = 15$			
		Welch	.049	.051	.051	.053	.074	.056	.045	.047	.039	.044	.042	.040

表 4　n = 20 时，双侧检验（经典的双样本 t 检验和 Welch 的 t 检验，每个检验都具有 α = 0.05）的模拟显著性水平

			基本总体：N(0,2) 第二个总体：N(0, σ_2)				基本总体：Chi(2) 第二个总体：卡方				基本总体：CN(.8,4) 第二个总体：CN(.8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$	方法	$n_1 = 20$				$n_1 = 20$				$n_1 = 20$			
10	.5	2T	.019	.052	.087	.115	.028	.048	.087	.119	.021	.048	.067	.079
		Welch	.050	.054	.053	.053	.044	.054	.061	.061	.048	.042	.035	.028
20	1.0	2T	.048	.049	.052	.053	.057	.046	.052	.056	.049	.044	.042	.040
		Welch	.045	.049	.051	.050	.055	.044	.050	.052	.047	.042	.040	.037
30	1.5	2T	.086	.054	.039	.032	.098	.047	.035	.033	.075	.047	.033	.022
		Welch	.054	.054	.053	.052	.068	.047	.051	.053	.041	.043	.044	.042
40	2.0	2T	.107	.049	.026	.016	.123	.046	.027	.019	.107	.047	.026	.016
		Welch	.048	.049	.046	.047	.070	.054	.046	.045	.044	.043	.043	.042

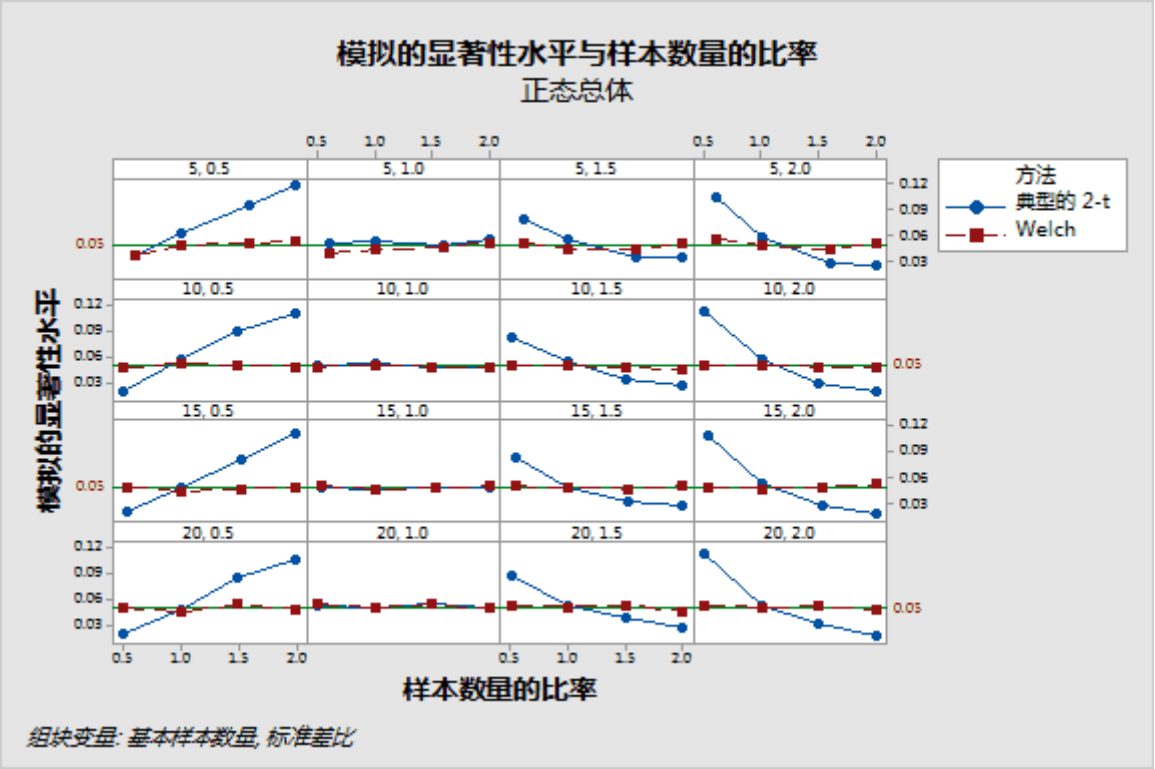

图 1　双侧检验（经典的双样本 t 检验和 Welch 的 t 检验，每个检验都具有 $\alpha = 0.05$）的模拟显著性水平，基于从两个具有等方差或不等方差（根据样本数量比率绘制）的正态总体生成的样本对。

模拟结果显示，对于相对较小的样本数量，经典的双样本 t 检验相对于非正态性检验更稳健，但对等方差假设敏感，除非双样本设计接近平衡。这是图 1、2 和 3 中的图形显示。基于经典的双样本 t 检验的模拟显著性水平曲线在样本数量比率为 1.0 所在的点处与目标水平线相交，即使方差差异非常大也是如此。对于所有三个分布族（正态、卡方和污染正态总体），如果样本数量不同，则仅在方差相等时，经典的双样本 t 检验的模拟显著性水平接近于目标水平。这已在图 1、图 2 和图 3 中的控制图的第二列中标明。

在设计不平衡并且方差不相等时，经典的 t 检验性能反而不良。即使方差之间出现小的不一致也会带来很多问题。对于那些不等方差、不平衡设计，数据的正态性不会提高模拟显著性水平。实际上，随着样本数量的增大，模拟显著性水平越来越远离目标水平，这与父总体无关。在从具有更大方差的总体中抽取更大样本时，模拟显著性水平小于目标水平。在从具有更小方差的总体中抽取更大样本时，模拟显著性水平大于目标水平。在不等方差假设条件下检查经典的双样本 t 检验统计量的渐近分布时，Arnold（1990，第 372 页）作出了类似的评论。

另一方面，Welch 双样本 t 检验不受偏离等方差假设影响，如图 1、图 2 和图 3 中所示。由于 Welch t 检验不是在等方差假设条件下推导出的，因此这不足为奇。推导出 Welch 的 t 检验所依据的正态假设似乎仅在这两个样本数量的最小值非常小时才显得比较重要。但对于较大样本，此检验不受偏离正态性假设的影响。这已在图 2 和图 3 中标明，其中，在两个样本数量的最小值为 15 时，模拟显著性水平持续接近目标水平。当这两个样本源自自由度为 2 的卡方分布，并且这两个样本数量均为 15 时，模拟显著性水平为 0.042（请参见表 3）。

在两个样本数量的最小值足够大时，离群值似乎不应影响 Welch 的 t 检验的性能。表 3 和图 3 显示，在两个样本数量的最小值至少为 15 时，则模拟显著性水平接近目标水平（在标

准差比值分别为 0.5、1.0、1.5 和 2.0 时，模拟显著性水平为 0.045、0.045、0.041、0.037）。

这些结果表明，在大多数实际应用中，Welch 双样本 t 检验在模拟显著性水平或 I 类错误率方面的性能高于经典的双样本 t 检验。

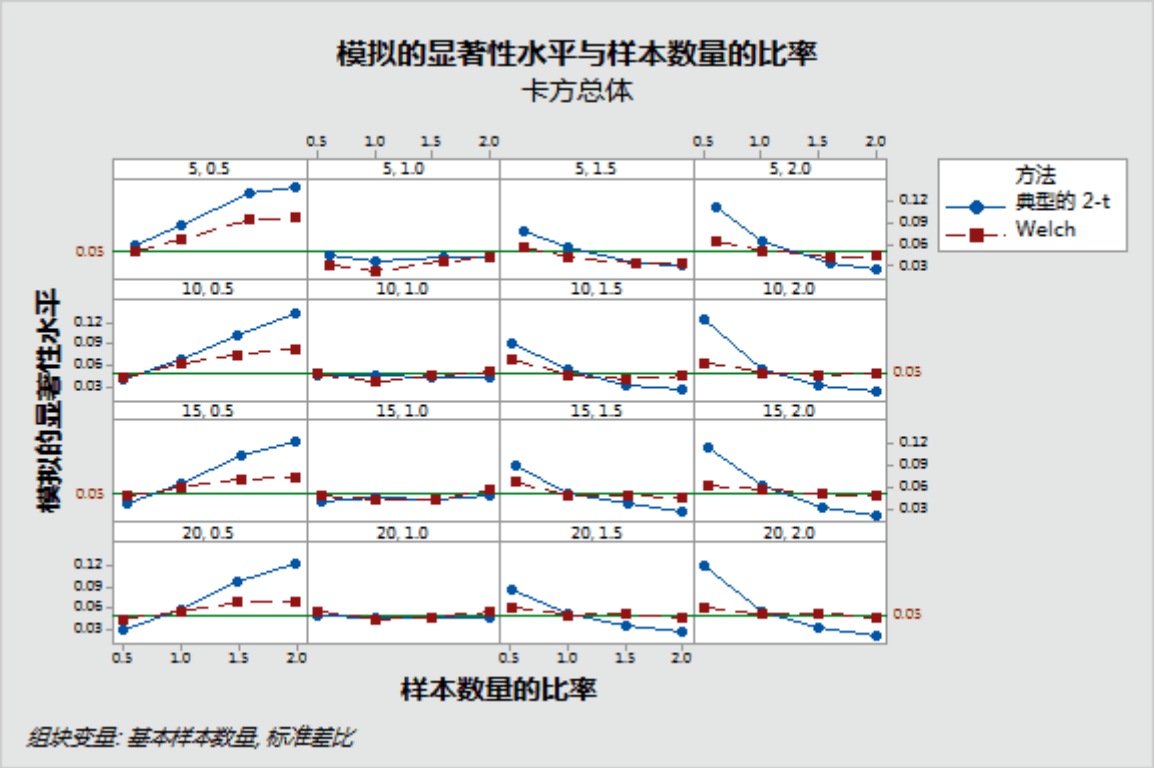

图 2　双侧检验（经典的双样本 t 检验和 Welch t 检验）的模拟显著性水平，基于从两个具有等方差或不等方差（根据样本数量比率绘制）的正态总体生成的样本对。

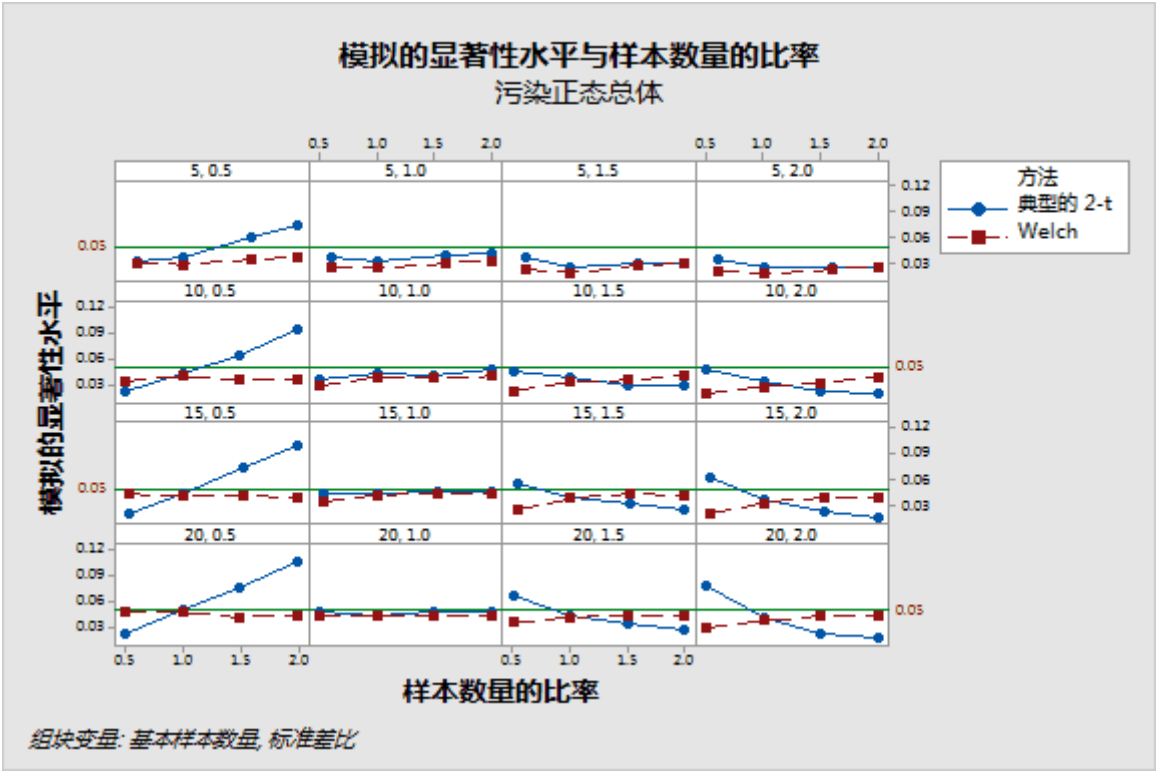

图 3　双侧检验（经典的双样本 t 检验和 Welch t 检验）的模拟显著性水平，基于从两个具有等方差或不等方差（根据样本数量比率绘制）的正态总体生成的样本对。

附录 B：这两个检验的功效函数比较

我们想要确定在哪些条件下 Welch 的 t 检验的功效函数等于或大致等于经典的双样本 t 检验的功效函数。

通常，t 检验（单样本或双样本）的功效函数已知，并在许多总体中讨论过（Pearson 和 Hartley，1952；Neyman 等人，1935；Srivastava，1958）。以下定理阐述了双样本设计中三个不同备择假设中的各个假设的功效函数。

定理 B1

在正态性和方差齐性的假设条件下，具有名义数量 α 的双侧双样本 t 检验的功效函数可能表示为样本数量的函数，差值 $\delta = \mu_1 - \mu_2$ 表示为

$$\pi(n_1, n_2, \delta) = 1 - F_{d_C,\lambda}\left(t_{d_C}^{\alpha/2}\right) + F_{d_C,\lambda}\left(-t_{d_C}^{\alpha/2}\right)$$

其中，$F_{d_C,\lambda}(.)$ 是具有 $d_C = n_1 + n_2 - 2$ 自由度和非中心参数的非中心 t 分布的 C.D.F

$$\lambda = \frac{\delta}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

此外，与备择假设 $\mu_1 > \mu_2$ 关联的功效函数给定为

$$\pi(n_1, n_2, \delta) = 1 - F_{d_C,\lambda}\left(t_{d_C}^{\alpha}\right)$$

另一方面，在检验备择假设 $\mu_1 < \mu_2$ 时，功效表示为

$$\pi(n_1, n_2, \delta) = F_{d_C,\lambda}\left(-t_{d_C}^{\alpha}\right)$$

虽然上述定理的结果已知，但基于 Welch 改进的 t 检验的检验功效函数没有专门的参考资料。可根据单因子方差分析模型推导得出的近似功效函数断定近似情况（请参见 Kulinskaya 等人，2003）。不巧的是，此功效函数仅适用于双侧备择假设。但是，双样本设计其实是这样一种特殊情况，我们可以采用不同方法来获取这三个备择假设中各个假设的 Welch 的 t 检验的（精确）功效函数。这些函数已在以下定理中给定。

定理 B2

在总体呈正态分布（但不一定具有相同的方差）的假设条件下，双侧 Welch 的 t 检验（具有名义数量 α）的功效函数可以表示为样本数量的函数，差值 $\delta = \mu_1 - \mu_2$ 表示为

$$\pi_W(n_1, n_2, \delta) = 1 - G_{d_W,\lambda_W}\left(t_{d_W}^{\alpha/2}\right) + G_{d_W,\lambda_W}\left(-t_{d_W}^{\alpha/2}\right)$$

其中，$G_{d,\lambda}(.)$ 是非中心 t 分布的 C.D.F，其自由度 d_W 表示为

$$d_W = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1 - 1)} + \frac{\sigma_2^4}{n_2^2(n_2 - 1)}}$$

并且具有非中心参数

$$\lambda_W = \frac{\delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

对于单侧备择检验，功效函数给定为

$$\pi_W(n_1,n_2,\delta) = 1 - G_{d_W,\lambda_W}\left(t^{\alpha}_{d_W}\right)$$

和

$$\pi_W(n_1,n_2,\delta) = G_{d_W,\lambda_W}\left(-t^{\alpha}_{d_W}\right)$$

以分别对备择假设 $\mu_1 > \mu_2$ 和备择假设 $\mu_1 < \mu_2$ 检验原假设。

附录 D 中给出了此结果的论证。

在我们比较这两个功效函数之前，请注意，由于经典的双样本 t 检验是在总体方差相等的另一个假设条件下推导出的，则在对 Welch 的 t 检验应用第二个假设时，应该比较这两个检验的理论功效函数。

理论上说，我们知道，在正态性和等方差假设条件下，

$\pi(n_1,n_2,\delta) \geq \pi_W(n_1,n_2,\delta)$ 适用于所有 n_1,n_2,δ

下一结果说明了两个函数（近似）相等的条件。

定理 B3

在正态性和方差齐性假设条件下，我们可以得出以下结论：

1. 如果 $n_1 \sim n_2$，则 $\pi(n_1,n_2,\delta) \sim \pi_W(n_1,n_2,\delta)$ 对每个差值 δ 都成立。尤其是，如果 $n_1 = n_2$，则 $\pi(n_1,n_2,\delta) = \pi_W(n_1,n_2,\delta)$ 对每个差值 δ 都成立，因此可知 Welch 的 t 检验的功效与经典的双样本 t 检验是一样的。
2. 如果 n_1 和 n_2 比较小，并且 $n_1 \neq n_2$，则 Welch 的 t 检验的功效小于经典的双样本 t 检验。但是，如果 n_1 和 n_2 比较大，则 $\pi(n_1,n_2,\delta) \sim \pi_W(n_1,n_2,\delta)$（与样本数量之间的差值无关）。

附录 E 中提供了此结果的论证。

在方差齐性假设条件下，两个检验的功效函数中的非中心参数是相同的。功效函数之间的差值只依赖于各自自由度之间的差值。从理论上讲，我们知道，在所述的假设情况下，经典的 t 检验为 UMP（一致最优势检验），因此它具有更高的自由度。但上述结果表明，如果设计平衡或近似平衡，则功效函数相同或近似相同。经典的 t 检验的功效明显高于 Welch 的 t 检验的唯一情况就是设计明显不平衡，并且样本较小。不幸的是，附录 A 中已经显示了经典的双样本 t 检验对于等方差假设特别敏感，现在也会出现这种情况。因此，在实际应用中，Welch 的 t 检验功效函数更可靠。

我们通过以下示例解释了定理 B3 的结果，该示例中，两个正态总体具有相同的标准差 3。基于定理 B1 和定理 B2 的（双侧）功效函数的功效值根据以下四种情况计算得到：

1. 两个样本都很小，但数量相同（$n_1 = n_2 = 10$）。
2. 两个样本都很小，但一个样本是另一个样本的两倍（ $n_1 = 10, n_2 = 20$）。
3. 一个样本很小，另一个样本数量中等，但中等样本的数量是小样本的四倍（$n_1 = 10, n_2 = 40$）。
4. 一个样本数量中等，另一个样本很大，但大样本的数量是中等样本的四倍（$n_1 = 50, n_2 = 200$）。

假设这两个检验的 $\alpha = 0.05$，则会按差值 $\delta = 0.0, 0.5, 1.0, 1.5, 2.0, \ldots 5.0$ 计算出每种情况下的功效函数。结果将显示在表 5 中，并且在图 4 中绘制这些函数。

表 5　双侧经典的双样本 t 检验和双侧 Welch t 检验（$\alpha = 0.05$）的理论功效函数比较。样本数量 n_1 和 n_2 固定，将按差值 δ（范围从 0.0 到 5.0）计算功效函数。

δ	0.0	0.5	1.0	1.5	2.0	2.5	3	3.5	4	4.5	5.0
$n_1 = n_2 = 10$											
$\pi(n_1, n_2, \delta)$	.05	.064	.109	.185	.292	.422	.562	.694	.805	.887	.941
$\pi_W(n_1, n_2, \delta)$	.05	.064	.109	.185	.292	.422	.562	.694	.805	.887	.941
$n_1 = 10, n_2 = 20$											
$\pi(n_1, n_2, \delta)$	.05	.070	.132	.239	.383	.547	.703	.828	.913	.962	.986
$\pi_W(n_1, n_2, \delta)$	.05	.070	.129	.231	.371	.531	.686	.813	.902	.955	.982
$n_1 = 10, n_2 = 40$											
$\pi(n_1, n_2, \delta)$	.05	.075	.152	.283	.455	.637	.791	.899	.959	.986	.996
$\pi_W(n_1, n_2, \delta)$	.05	.072	.142	.261	.419	.592	.748	.865	.938	.976	.992
$n_1 = 50, n_2 = 200$											
$\pi(n_1, n_2, \delta)$	.05	.182	.556	.883	.987	.999	1.	1.	1.	1.	1.
$\pi_W(n_1, n_2, \delta)$	.05	.180	.548	.877	.986	.999	1.	1.	1.	1.	1.

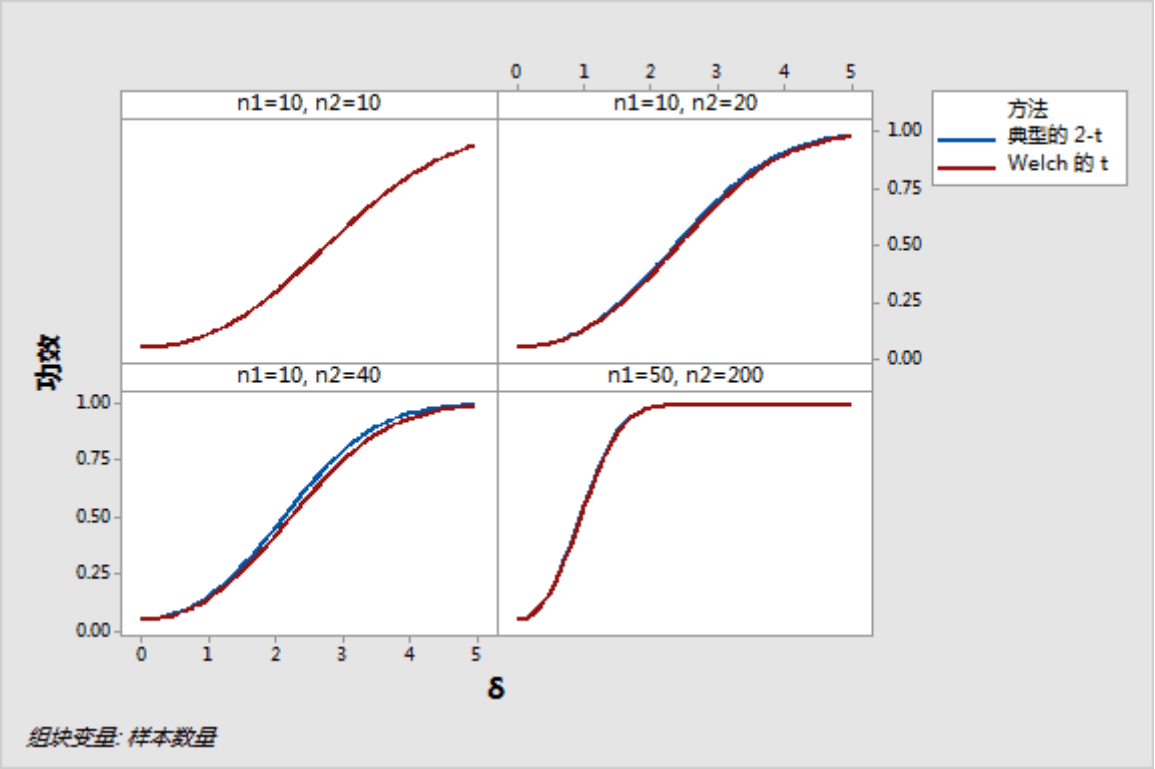

图 4　双侧经典的双样本 t 检验和双侧 Welch 的 t 检验的理论功效函数与均值差δ值的关系图。这两种检验都使用 $\alpha = 0.05$。假设总体呈正态分布，并且具有相同的标准差 3。

模拟研究 B

此模拟研究的目的是将与经典的双样本 t 检验的功效水平和与平衡设计（方差假定为不等）中 Welch 双样本 t 检验的功效水平进行比较。这些研究中的试验类似于附录 A 中讨论的试验。

在第一组试验中，我们根据具有不等方差的正态总体生成了等数量的样本对。基本总体固定为 $N(0,2)$，并且选择了第二个正态总体，使得标准差比值 $\rho = \sigma_2/\sigma_1$ 等于 0.5、1.5 和 2。类似地，在第二组中，这两个样本是从方差不等的卡方分布（基本总体是 Chi(2)）中抽取的。在最后一组试验中，样本对是根据之前附录 A 中定义的污染正态分布（基本总体 CN(.8,4)）生成的。

在每组试验中，对于与样本数量 $n = n_1 = n_2 = 5, 10, 15, 20, 25, 30$ 关联进行的每次检验我们计算了模拟功效水平（按给定的可检测差值 δ）。在每次试验中，如果在原假设为假时拒绝原假设，则模拟功效水平计算为实例的比例。对于所有试验，我们将两个样本中的第一个样本当作基本总体，将均值差指定为基本总体的一个标准差。更特别地，我们确定 $\delta = 1.0 \times \sigma_1 = 2.0$，因为此差值对于此研究中的所有三个分布族而言相对比较小。表 6 中报告了模拟结果，并且这些结果显示在图 5、图 6 和图 7 中。

结果与汇总

表 5 和图 4 中的结果显示，在等方差假设条件下，如定理 B3 中所述，平衡设计中的理论功效函数相同。此外，在样本数量相对比较小但接近相同数量时，这两个函数会产生近似相等的功效值。功效函数之间开始出现一些显著差异仅在以下情况下出现：样本量相对比较小，一个样本量大约是另一个样本量的四倍（例如，在 $n_1 = 10, n_2 = 40$ 时）。即使在这种情况下，

基于经典的双样本 t 检验的理论功效值也仅略高于基于 Welch 的 t 检验的功效值。最终，当设计明显不平衡但样本数量（相对）较大时，这两个功效函数基本上相同，如定理 B3 中所述。

此外，在方差不等的平衡设计中，这两个检验可产生实际相同的功效值。但在非常小的样本（$n < 10$）中，经典的双样本 t 检验的执行效果略好。

表 6　方差不等的平衡设计中经典的双样本 t 检验和 Welch 的检验的模拟功效水平比较

		基本总体：N(0,2)			基本总体：Chi(2)			基本总体：CN(.8,4)		
n	$\frac{\sigma_2}{\sigma_1}$	.5	1.5	2.0	.5	1.5	2.0	.5	1.5	2.0
5	2T	0.431	0.196	0.152	0.555	0.281	0.215	0.579	0.373	0.335
	Welch	0.366	0.166	0.119	0.424	0.25	0.184	0.521	0.32	0.283
10	2T	0.77	0.385	0.27	0.846	0.438	0.324	0.79	0.51	0.435
	Welch	0.747	0.372	0.253	0.832	0.427	0.308	0.776	0.493	0.417
15	2T	0.916	0.539	0.387	0.948	0.565	0.424	0.898	0.615	0.508
	Welch	0.908	0.532	0.375	0.945	0.557	0.413	0.891	0.605	0.497
20	2T	0.971	0.682	0.497	0.982	0.68	0.521	0.952	0.702	0.573
	Welch	0.969	0.677	0.487	0.981	0.676	0.511	0.947	0.697	0.563
25	2T	0.99	0.779	0.591	0.994	0.765	0.605	0.98	0.783	0.641
	Welch	0.99	0.777	0.582	0.994	0.762	0.597	0.979	0.778	0.636
30	2T	0.998	0.851	0.675	0.998	0.826	0.676	0.994	0.839	0.699
	Welch	0.998	0.849	0.67	0.998	0.824	0.668	0.994	0.836	0.694

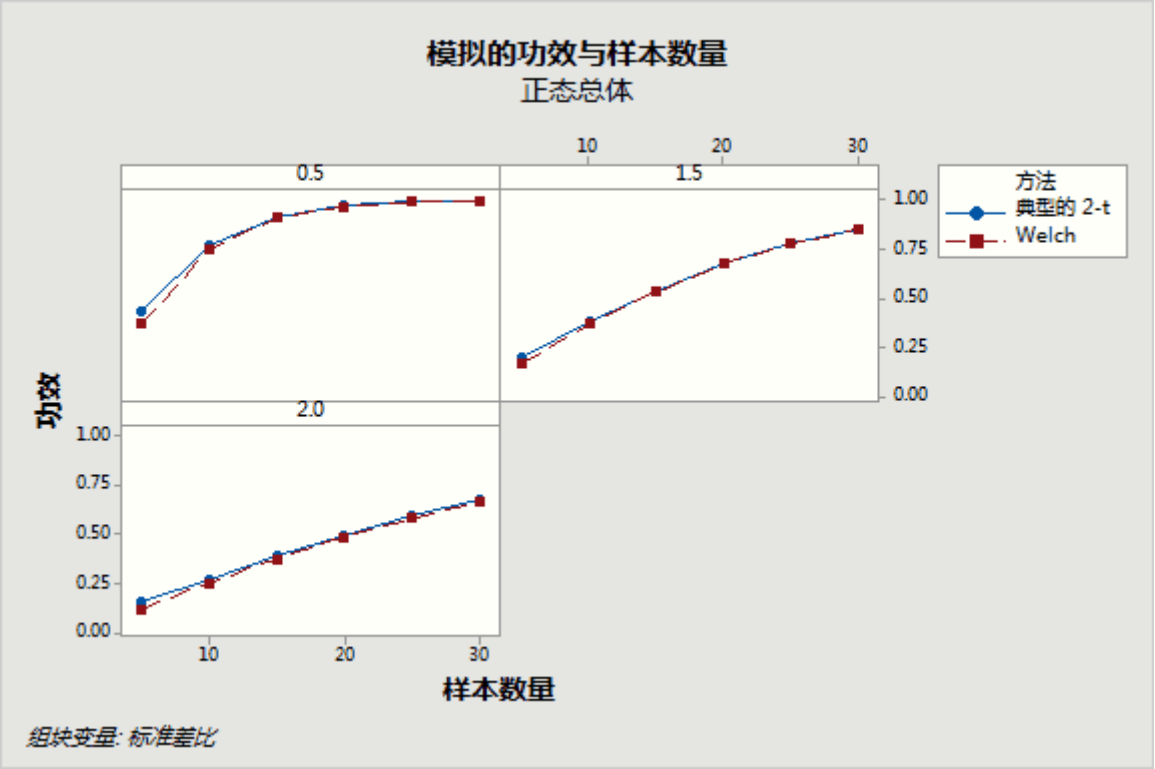

图 5　方差不等的平衡设计中经典的双样本 t 检验和 Welch 的双样本 t 检验的模拟功效水平比较。样本是从方差不等的正态总体中抽取的，其标准差比值为 0.5、1.5 和 2.0。

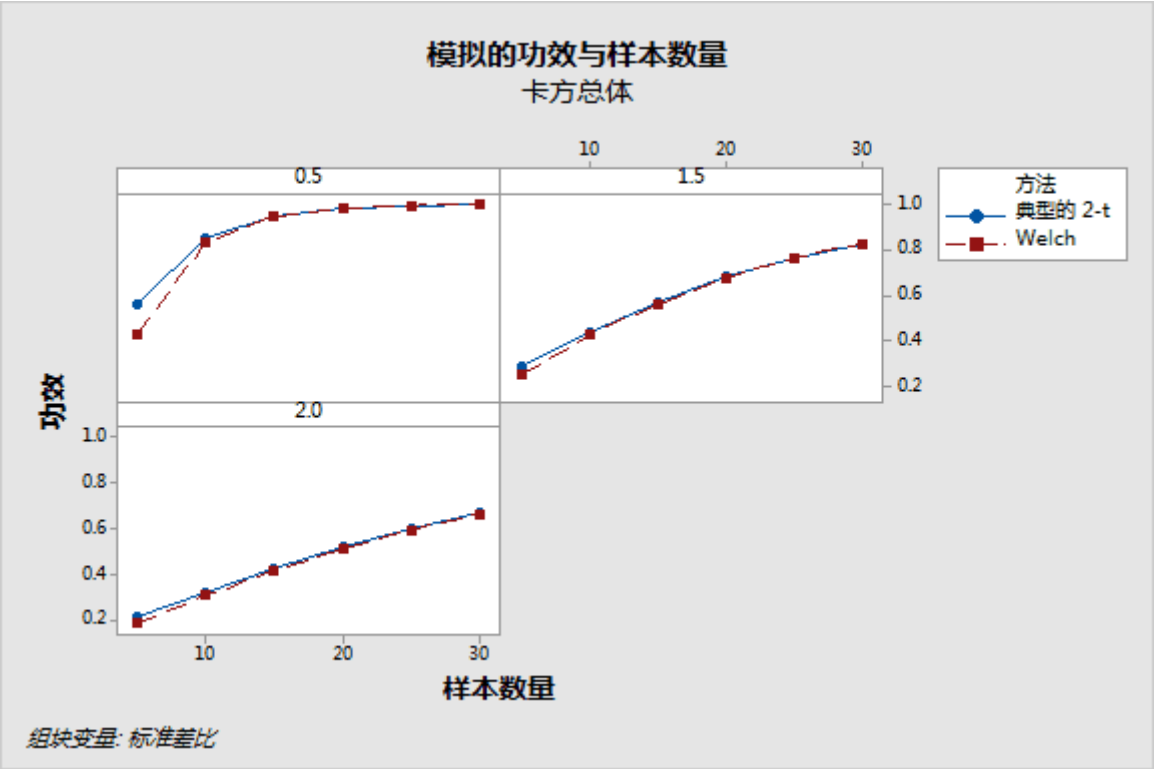

图 6　方差不等的平衡设计中经典的双样本 t 检验和 Welch 的双样本 t 检验的模拟功效水平比较。样本是从方差不等的卡方总体中抽取的，其标准差比值为 0.5、1.5 和 2.0。

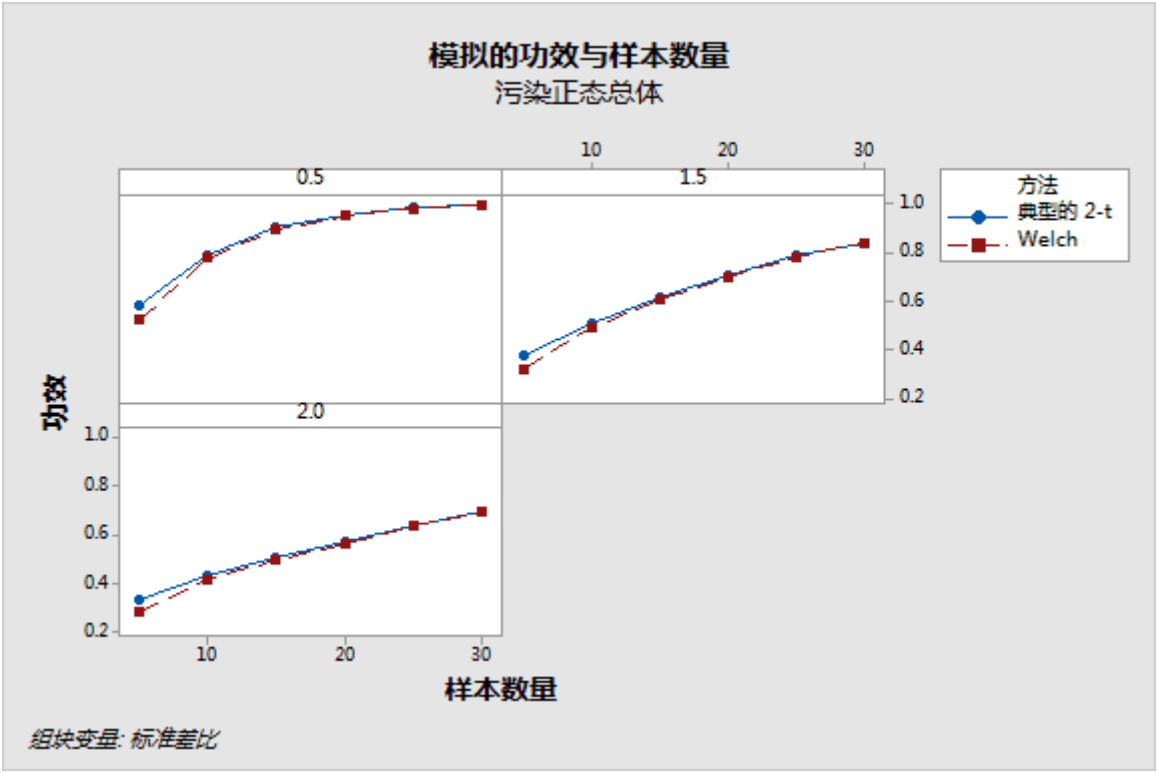

图 7　方差不等的平衡设计中经典的双样本 t 检验和 Welch 的双样本 t 检验的模拟功效水平比较。样本是从方差不等的污染正态总体中抽取的，其标准差比值为 0.5、1.5 和 2.0。

附录 C：功效和样本数量以及对正态性的敏感度

在“协助”中，用于比较两个总体的均值的功效分析基于 Welch 的 t 检验的功效函数。如果此函数对推导出它所依据的正态假设敏感，则功效分析可能会产生错误的结论。鉴于此原因，我们进行了模拟研究，以检查此函数对正态假设的敏感度。评估非正态性的敏感度其实就是要比较模拟功效水平和根据理论功效函数计算的功效水平之间的一致性，这里样本是根据非正态分布生成的。正态分布可用作对照总体，因为按照定理 B2，在样本根据非正态总体生成时，其模拟功效水平和理论功效水平是很接近的。

模拟研究 C

此研究使用正态、卡方和污染正态这三种分布，分三个部分进行。有关详细信息，请参考附录 A。对于此研究的每个部分，对给定的样本数量 n_1 和 n_2 给定的可检测差值为 δ 时生成的数据实例中，模拟功效就是能够在原假设为假时拒绝原假设所占的比率。在所有情况下，要检测的差值按基本总体中的一个标准差指定。对于本研究中的所有三个分布族而言，该值为 $\delta = 1.0 \times \sigma_1 = 2.0$。为进行比较，还会计算基于 Welch 的 t 检验的理论功效值。

模拟结果和汇总

结果显示，对于相对较小的样本数量，Welch t 检验的功效函数相对于正态性假设更稳健。通常，在两个样本数量的最小值达到 15 时，模拟功效值接近其对应的目标理论功效水平（请参见表 7-10 和图 8-10）。

表 7-10 显示了双侧 Welch t 检验（$\alpha = 0.05$）的模拟功效水平，基于根据正态总体、偏斜总体（卡方）和污染正态总体生成的样本对。这些样本对来自相同的分布族，但父总体的方差不一定相等。为进行比较，还计算了理论功效值。

表 7　n=5 时，双侧 Welch t 检验（$\alpha = 0.05$）的模拟功效水平

			基本总体：N(0,2)				基本总体：Chi(2)				基本总体：CN(.8,4)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$		$n_1 = 5$				$n_1 = 5$				$n_1 = 5$			
3	.6	观测值	.288	.158	.113	.091	.432	.305	.211	.149	.361	.257	.234	.220
		目标值	.353	.192	.116	.092	.353	.192	.116	.092	.353	.192	.116	.092
5	1.0	观测值	.370	.252	.169	.121	.427	.334	.248	.189	.522	.380	.319	.284

			基本总体：N(0,2)				基本总体：Chi(2)				基本总体：CN(.8,4)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$		$n_1 = 5$				$n_1 = 5$				$n_1 = 5$			
		目标值	.389	.286	.190	.137	.389	.286	.190	.137	.389	.286	.190	.137
8	1.6	观测值	.387	.326	.242	.179	.427	.364	.286	.225	.573	.453	.374	.319
		目标值	.400	.345	.260	.193	.400	.345	.260	.193	.400	.345	.260	.193
10	2.0	观测值	.390	.351	.272	.208	.421	.373	.296	.235	.590	.483	.394	.336
		目标值	.402	.364	.291	.223	.402	.364	.291	.223	.402	.364	.291	.223

表 8 n=10 时，双侧 Welch t 检验（$\alpha = 0.05$）的模拟功效水平

			基本总体：N(0,2)				基本总体：Chi(2)				基本总体：CN(.8,4)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$		$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
5	.5	观测值	.651	.346	.197	.131	.768	.493	.320	.221	.689	.484	.404	.358
		目标值	.666	.364	.206	.139	.666	.364	.206	.139	.666	.364	.206	.139
10	1.0	观测值	.742	.556	.369	.254	.831	.612	.430	.308	.776	.619	.496	.419
		目标值	.745	.562	.337	.259	.745	.562	.337	.259	.745	.562	.337	.259
15	1.5	观测值	.765	.641	.483	.358	.865	.679	.511	.377	.792	.679	.547	.456
		目标值	.767	.643	.483	.352	.767	.643	.483	.352	.767	.643	.483	.352
20	2	观测值	.774	.683	.549	.417	.898	.737	.565	.448	.797	.716	.596	.490
		目标值	.777	.686	.551	.422	.777	.686	.551	.422	.777	.686	.551	.422

表 9　n=15 时，双侧 Welch t 检验（$\alpha = 0.05$）的模拟功效水平

			基本总体：N(0,2)				基本总体：Chi(2)				基本总体：CN(.8,4)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$		$n_1 = 15$				$n_1 = 15$				$n_1 = 15$			
8	.53	观测值	.857	.569	.342	.229	.871	.651	.421	.293	.853	.632	.505	.428
		目标值	.861	.568	.338	.221	.861	.568	.338	.221	.861	.568	.338	.221
15	1.0	观测值	.906	.745	.535	.368	.942	.763	.563	.415	.891	.760	.611	.500
		目标值	.910	.753	.541	.379	.910	.753	.541	.379	.910	.753	.541	.379
23	1.53	观测值	.928	.831	.667	.502	.975	.858	.676	.517	.898	.825	.698	.572
		目标值	.925	.830	.670	.509	.925	.830	.670	.509	.925	.830	.670	.509
30	2.0	观测值	.933	.861	.737	.589	.984	.903	.750	.598	.902	.847	.742	.619
		目标值	.931	.863	.736	.589	.931	.863	.736	.589	.931	.863	.736	.589

表 10　n=20 时，双侧 Welch t 检验（$\alpha = 0.05$）的模拟功效水平

			基本总体：N(0,2)				基本总体：Chi(2)				基本总体：CN(.8,4)			
		$\frac{\sigma_2}{\sigma_1}$	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0	.5	1.0	1.5	2.0
n_2	$\frac{n_2}{n_1}$		$n_1 = 20$				$n_1 = 20$				$n_1 = 20$			
10	.5	观测值	.938	.687	.426	.275	.920	.698	.486	.333	.923	.716	.568	.476
		目标值	.941	.686	.424	.277	.941	.686	.424	.277	.941	.686	.424	.277
20	1.0	观测值	.971	.866	.672	.485	.981	.858	.670	.506	.952	.856	.696	.567
		目标值	.971	.869	.673	.489	.971	.869	.673	.489	.971	.869	.673	.489

30	1.5	观测值	.977	.923	.791	.629	.995	.932	.785	.631	.960	.908	.798	.662
		目标值	.978	.922	.791	.628	.978	.922	.791	.628	.978	.922	.791	.628
40	2.0	观测值	.983	.950	.858	.724	.998	.966	.864	.726	.958	.929	.845	.725
		目标值	.981	.945	.854	.719	.981	.945	.854	.719	.981	.945	.854	.719

如果这两个样本是根据正态总体生成的，则模拟功效值与理论功效值一致，即使是非常小的样本也如此。如图 7 所示，理论和模拟功效曲线实际上不可区分。这些结果与定理 B2 一致。

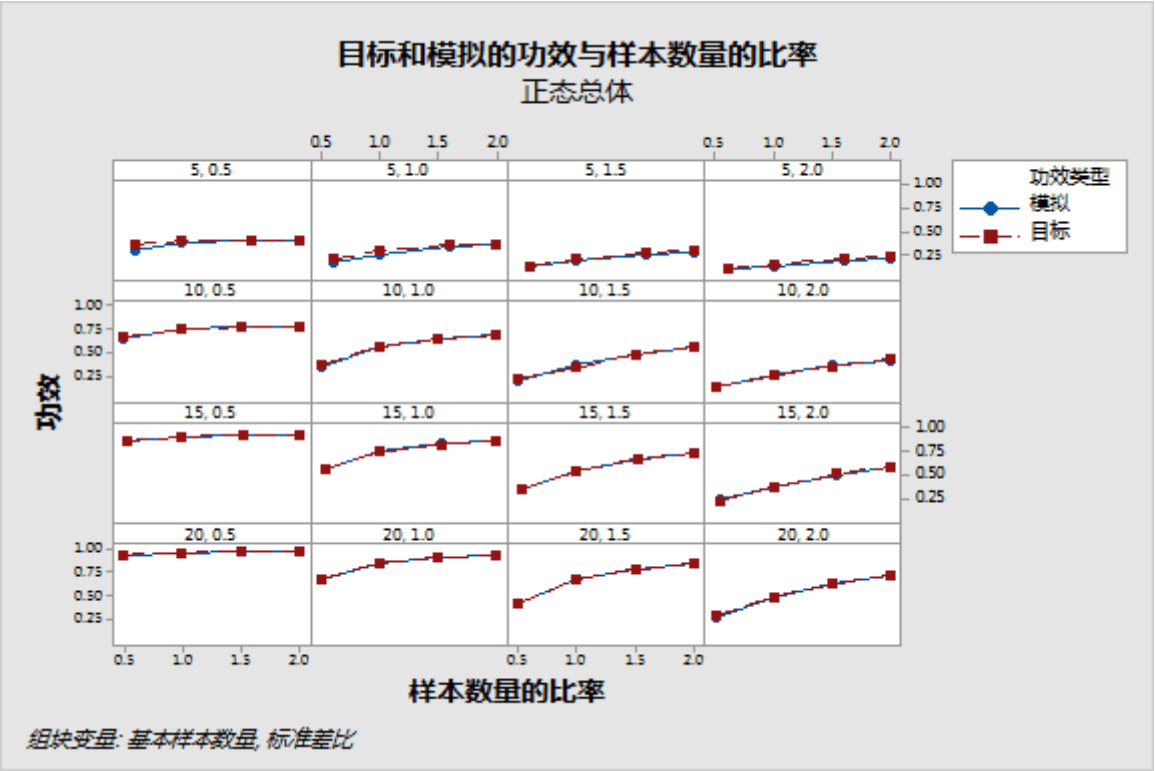

图 8　双侧 Welch t 检验（$\alpha = 0.05$）的模拟和目标理论功效水平，基于根据两个正态总体生成的样本对以及根据样本数量比值绘制的等方差或不等方差。

如果样本根据偏斜卡方分布生成，则对于非常小的样本，模拟功效值高于理论功效值，但随着样本数量的增大，这两个功效值更接近。图 9 显示，在这两个样本数量的最小值至少为 10 时，目标理论和模拟功效曲线一致保持接近。这说明，在样本数量相对较小时，偏斜数据对 Welch t 检验的功效函数没有显著影响。

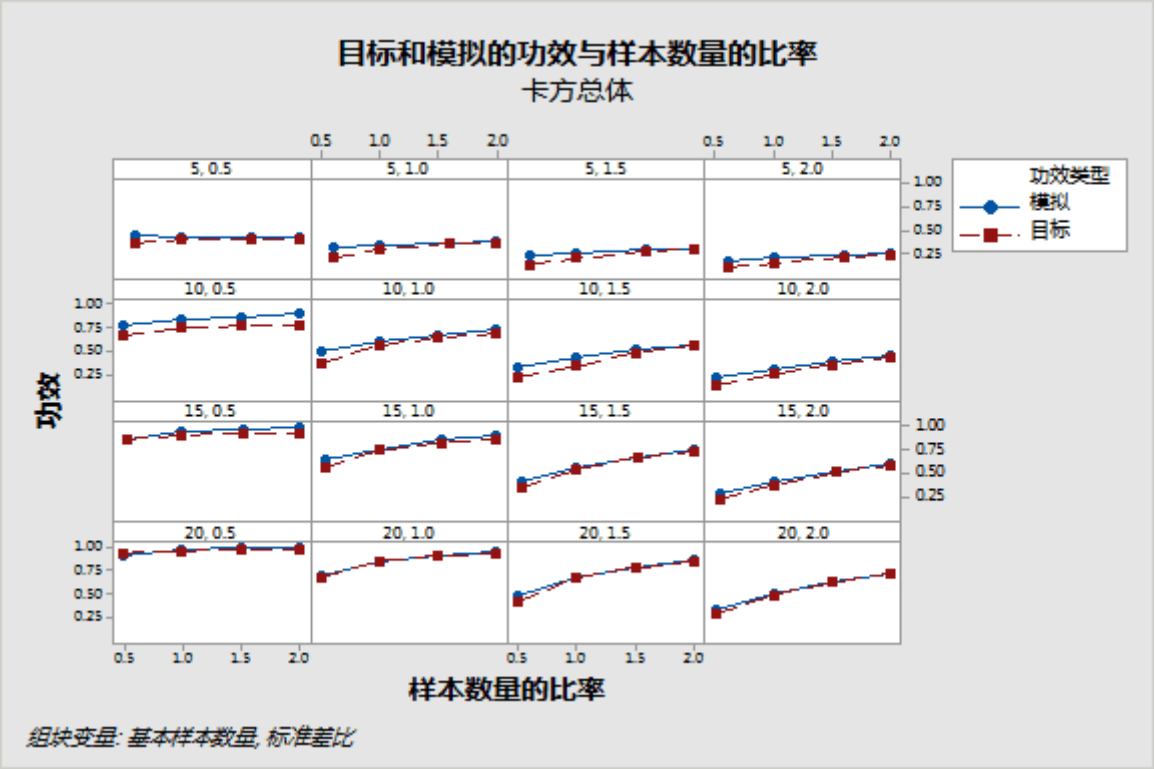

图 9　双侧 Welch t 检验（$\alpha = 0.05$）的模拟和目标理论功效水平，基于根据两个正态总体生成的样本对以及根据样本数量比值绘制的等方差或不等方差。

此外，仅在样本数量非常小时，离群值才可能对功效函数产生影响。通常，在出现离群值时，模拟功效值往往略大于目标理论功效值。如图 10 中所示，模拟和理论功效曲线并不十分接近，除非最小样本数量达到 15。

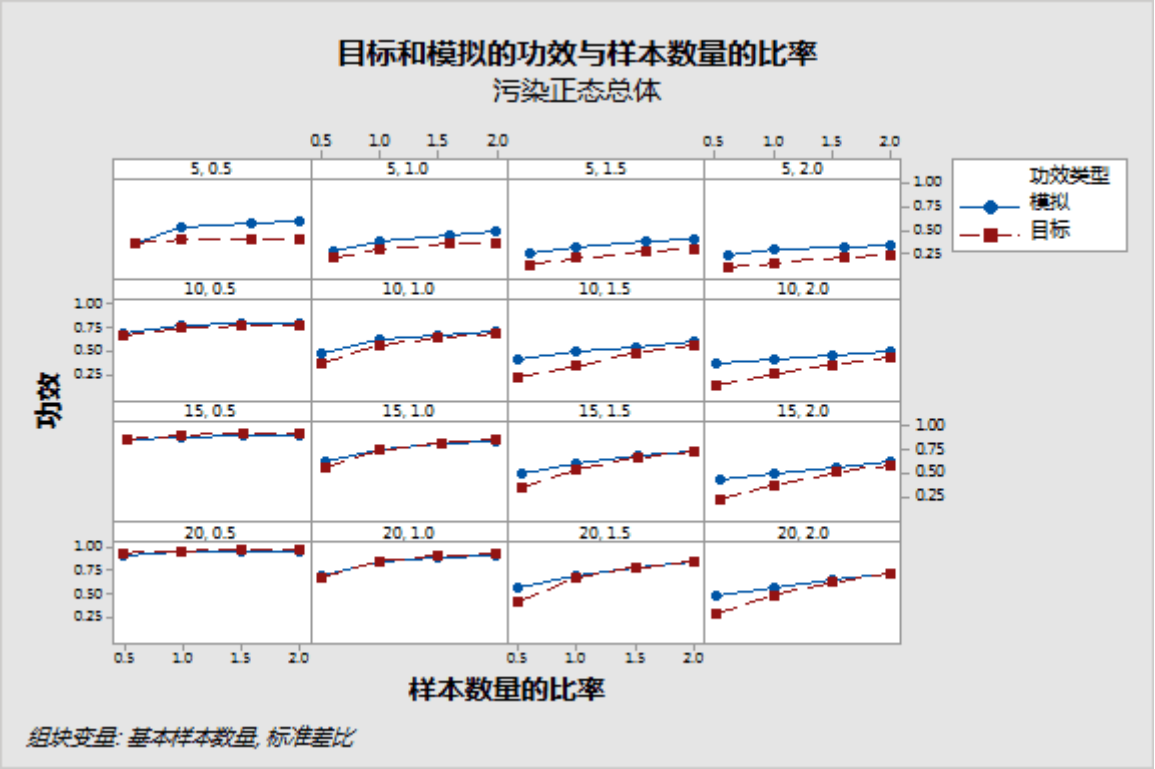

图 10　双侧 Welch t 检验（$\alpha = 0.05$）的模拟和目标理论功效水平，基于根据两个正态总体生成的样本对以及根据样本数量比值绘制的等方差或不等方差。

附录 D：定理 B2 的论证

对于双样本模型，用于在原假设条件下推导出检验统计量分布的 Welch 方法

$$t_w(x,y)=\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

基于以下分布的近似

$$V=\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}$$

与卡方分布成比例。具体地说，

$$\frac{d_W V}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

近似分布为卡方分布，自由度为 d_W，其中

$$d_W=\frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1-1)}+\frac{\sigma_2^4}{n_2^2(n_2-1)}}$$

（注，在单样本设置中，这可以简化为已知的经典结果 $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$）

考虑假设检验问题，其原假设 $H_A: \mu_1 \neq \mu_2$（或等效 $\delta \neq 0$），其备择假设 $H_o: \mu_1 = \mu_2$（或等效 $\delta = 0$）

在原假设条件下，功效函数

$$\pi(n_1,n_2,\delta)=\pi(n_1,n_2,0)=1-\Pr\left(-t_{d_W}^{\alpha/2}\leq\frac{\bar{x}-\bar{y}}{\sqrt{V}}\leq t_{d_W}^{\alpha/2}\right)\approx\alpha$$

其中，t_d^α 表示 t 分布的 $100\,\alpha$ 上百分位点，自由度为 d。

在备择假设条件下，

$$\frac{\bar{x}-\bar{y}}{\sqrt{V}}=\frac{\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}+\frac{\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}}{\sqrt{\frac{d_W V}{d_W\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}}}$$

具有近似非中心 t 分布，自由度为 d_W，并且具有非中心参数

$$\lambda_W=\frac{\delta}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$$

如前所述，

$$\frac{d_W V}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

近似分布为卡方分布，自由度为 d_W，并且

$$\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

分布为标准正态分布。

在备择假设条件下，它将有下列结果

$$\pi(n_1,n_2,\delta)=1-\Pr\left(-t_{d_W}^{\alpha/2}\le\frac{\bar{x}-\bar{y}}{\sqrt{V}}\le t_{d_W}^{\alpha/2}\right)\approx 1-G_{d_W,\lambda_W}\left(t_{d_W}^{\alpha/2}\right)+G_{d_W,\lambda_W}\left(-t_{d_W}^{\alpha/2}\right)$$

其中，$G_{d_W,\lambda}(.)$ 是自由度为 d_W，非中心参数为 λ（如上给定）的非中心 t 分布的 C.D.F。

附录 E：定理 B3 的论证

首先，请注意 d_W 可以重写为

$$d_W = \frac{\left(\frac{1}{n_1} + \frac{\rho^2}{n_2}\right)^2}{\frac{1}{n_1^2(n_1-1)} + \frac{\rho^4}{n_2^2(n_2-1)}}$$

其中，$\rho = \sigma_1/\sigma_2$。

类似地，与 Welch t 检验功效函数关联的非中心参数也可以写为

$$\lambda_W = \frac{\delta/\sigma_1}{\sqrt{1/n_1 + \rho^2/n_2}}$$

在等方差假设条件下，与经典的双样本 t 检验和 Welch 检验的功效函数关联的非中心参数一致。即

$$\lambda = \lambda_W = \frac{\delta}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

其中，σ 是双总体的共同方差。因此，这两个检验的功效函数之间的差别仅在于其各自自由度之间的差别。但在等方差假设条件下，与 Welch t 检验功效函数关联的自由度变成

$$d_W = \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^2}{\frac{1}{n_1^2(n_1-1)} + \frac{1}{n_2^2(n_2-1)}} = \frac{(n_1+n_2)^2(n_1-1)(n_2-1)}{n_1^2(n_1-1) + n_2^2(n_2-1)}$$

按照定理 1，与经典的双样本 t 检验的功效函数相关的自由度为 $d_C = n_1 + n_2 - 2$。在经过一些代数运算之后，我们得到

$$d_C - d_W = \frac{(n_1-n_2)^2(n_1+n_2-1)^2}{n_1^2(n_1-1) + n_2^2(n_2-1)} \geq 0$$

$d - d_W \geq 0$ 不足为奇，因为我们知道，在方差齐性假设条件下，经典的双样本 t 检验为 UMP（一致最优势检验），因此与其功效函数关联的自由度应会更高。

现在，如果 $n_1 \sim n_2$，则 $d \sim d_W$，因此，功效函数具有相同的量值顺序。尤其是，如果 $n_1 = n_2$，则这两个检验的功效函数相同。这证明了定理 B3 的第一部分。

如果 $n_1 \neq n_2$，则 $d_C - d_W > 0$，因此，Welch t 检验的功效比经典的双样本 t 检验低。

此外，如果样本数量较大，即，如果 $n_1 \to \infty$，并且 $n_2 \to \infty$，则 $d_C \to \infty$，并且 $d_W \to \infty$，因此与这两个检验关联的检验统计量的渐进分布是标准正态分布。因此，这些检验渐进等效，并会产生相同的渐进功效函数。