单样本 t 检验

概述

单样本 t 检验用于估计检验过程的平均值并将该平均值与目标值进行比较。该检验操作起来比较可靠，因为当样本大小适中时，它对正态性假设极不敏感。根据大多数统计教材中的内容，单样本 t 检验和平均值的 t 置信区间适合任何大小为 30 或以上的样本。

在本文中，我们介绍了对这个针对至少 30 个样本单位的一般规则进行评估的模拟方法。我们的模拟重点关注非正态性对单样本 t 检验产生的影响。我们也希望评估异常数据对检验结果的影响。

根据我们的研究，“协助”会自动对您的数据进行以下检查并在“报告卡”中显示研究结果：

- 异常数据
- 正态性（样本量是否足够大，因此正态性不是问题？）
- 样本量

有关单样本 t 检验方法的一般信息，请参见 Arnold (1990), Casella and Berger (1990), Moore and McCabe (1993), and Srivastava (1958)。

注意：本文中的研究结果也适用于“协助”中的配对 t 检验，因为配对 t 检验对配对差异样本应用单样本 t 检验方法。

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。当样本量较小时，异常数据会影响发现具有重要统计意义的结果的概率。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。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，单样本检验的“协助报告卡”会显示以下状态指标：

状态	条件
	没有异常数据点。
	至少有一个异常数据点，可能会影响检验结果。

正态性

单样本 t 检验根据总体呈正态分布的假设推导出来。幸运的是，当样本量足够大时，即使数据不呈正态分布，此方法也同样有效。

目标

我们想要确定非正态性对检验的 I 类和 II 类错误的影响，以提供有关样本量和正态性的指南。

方法

在进行单样本 t 检验或计算某一总体的平均值的 t 置信区间时，我们进行了模拟，以确定可以忽略正态性假设的样本量。

我们设计了第一项研究，以评估非正态性对检验的 I 类错误概率的影响。具体而言，我们希望推测出检验所需的对总体分布不敏感的最小样本量。我们对从正态和非正态总体中生成的小、中、大样本进行了单样本 t 检验。非正态总体包括轻度和重度偏态总体、对称轻尾和重尾总体，以及受污染的正态总体。正态总体用作比较的控制总体。对于每一种情况，我们计算并比较了模拟显著性水平与目标显著性水平（标准的显著性水平为 0.05）。如果检验方法很有效，则模拟显著性水平应接近 0.05。我们研究了所有不同条件下的模拟显著性水平，以评估它们始终接近目标水平的最小样本量，不管分布如何。有关详细信息，请参见附录 A。

在第二项研究中，我们研究了非正态性对检验的 II 类错误的影响。模拟的设计理念与第一项研究相同。不过，我们比较了不同条件下的模拟功效水平与使用单样本 t 检验的理论功效计算出的目标功效水平。有关详细信息，请参见附录 B。

结果

对于小至 20 的样本量，非正态性对检验的 I 类和 II 类错误概率的影响最小。然而，当样本的母体呈极度偏态分布时，可能需要较大的样本。在这些情况下，我们建议使用约 40 个样本。有关详细信息，请参见附录 A 和附录 B。

由于相对较小的样本检验很有效，因此“协助”不检验数据的正态性， 而是检查样本量并在报告卡中显示以下状态指标：

状态	条件
	样本量至少为 20，所以正态性不是问题。
	样本量小于 20，所以可能存在正态性问题。

样本量

通常情况下，假设检验进行收集证据，拒绝“无差异”的原假设。如果样本量太小，检验的功效可能不足以检测出平均值之间确实存在的差异，这将导致 II 类错误。因此，一定要保证样本量足够大，以高概率地检测到实际存在的重要差异。

目标

如果数据未提供足够的证据来否定原假设，我们希望确定样本量是否足够大，从而检验能够以较高的高概率检测到所需的实际差异。尽管样本量规划的目标是确保样本量足够大，能够以较高的高概率检测到重要差异，但它们也不应太大，使得无意义的差异成为具有重要统计意义的高概率差异。

方法

功效和样本量分析基于用于进行统计分析的特定检验的理论功效。如前所述，当样本量至少为 20 时，单样本 t 检验的功效对正态假设不敏感。功效取决于样本量、目标平均值与总体平均值之间的差异，以及总体的方差。有关详细信息，请参见附录 B。

结果

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，如果用户提供所需的特定实际差异，“协助”将计算差异检测概率为 80% 和 90% 的样本量。

没有报告一般结果，因为这些结果取决于用户的具体实例。但是，您可以参阅附录 B，了解有关单样本 t 检验的功效的详细信息。

检查功效和样本量时，单样本 t 检验的“协助报告卡”会显示以下状态指标：

状态	条件
	检验发现平均值与目标值之间存在差异，所以功效不是问题。 或 功效是足够的。检验未发现平均值与目标值之间存在差异，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。检验未发现平均值与目标值之间存在差异，但样本量足够大，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。
	功效可能不够。检验未发现平均值与目标值之间存在差异，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。检验未发现平均值与目标值之间存在差异，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。
	检验未发现平均值与目标值之间存在差异。您没有指定平均值与目标值之间要检测的实际差异；因此，该报告将根据样本量、标准差和 alpha 值指出检测概率为 80% 和 90% 的差异。

参考书

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

Casella, G., & Berger, R. L. (1990). *Statistical inference.* Pacific Grove, CA: Wadsworth, Inc.

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.*

Moore, D.S. & McCabe, G.P. (1993). *Introduction to the practice of statistics*, 2nd ed. New York, NY: W. H. Freeman and Company.

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., & Hartley, H.O. (Eds.). (1954). *Biometrika tables for statisticians*, Vol. I. London: Cambridge University Press.

Srivastava, A. B. L. (1958). Effect of non-normality on the power function of t-test, *Biometrika, 45, 421-429.*

附录 A：非正态性对显著性水平的影响（检验的有效性）

在正态假设下，单样本 t 检验是一种一致最大功效（UMP）无偏 α 检验。也就是说，该检验的功效不与比其他任何与平均值相关的有偏 α 检验差。但是，当样本的母体不呈正态分布时，如果样本量足够大，则上述优化属性始终成立。换句话说，对于足够大的样本，单样本 t 检验的实际显著性水平约等于正态以及非正态数据的目标显著性水平，并且检验的功效也对正态假设不敏感（Srivastava, 1958）。

我们想要确定必须是多大的样本才被认为足够大，从而让 t 检验对正态假设不敏感。许多教材建议，如果样本量为 $n \geq 30$，则可以忽略正态假设，这是最切实可行的做法（Arnold, 1990; Casella & Berger, 1990; and Moore & McCabe, 1993）。这些附录中介绍的研究的目的是为了进行模拟研究，以通过检查不同的非正态分布对单样本 t 检验的影响，来评估这个一般规则。

模拟研究 A

我们要研究非正态性对检验的 I 类错误概率的影响，以评估不管分布如何，都能稳定地始终接近目标错误概率的最小样本量。

为此，我们使用从几个具有不同属性的分布中生成的各种大小（$n = 10, 15, 20, 25, 30, 35, 40, 50, 60, 80, 100$）的随机样本，执行了采用 $\alpha = 0.05$ 的双侧 t 检验。这些分布包括：

- 标准正态分布（N(0,1)）
- 对称和重尾分布，如自由度为 5 和 10 的 t 分布（t(5),t(10)）
- 位置为 0、标度为 1 的 Laplace 分布（Lpl）
- 用标度为 1 的指数分布（Exp）表示的偏态和重尾分布，自由度为 3、5 和 10 的卡方分布（Chi(3), Chi(5), Chi(10)）
- 对称和轻尾分布，如均匀分布（U(0,1)），两个参数设置为 3 的 Beta 分布（B(3,3)）
- 左偏和重尾分布（B (8,1)）

此外，为了评估异常值的直接影响，我们从受污染的正态分布中生成了样本，定义如下：

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

其中 p 被定义为混合参数，$1 - \mathrm{p}$ 是污染的比例或异常值的比例。我们为本次研究选择了两个受污染的正态总体：$\mathrm{CN}(0.9,3)$（10% 的总体成员是异常值），$\mathrm{CN}(.8,3)$（20% 的总体成员是异常值）。由于异常值，这是两个带长尾的对称分布。

对于每个样本量，我们从每个总体中提取了 10,000 个重复样本，并根据原假设 $\mu = \mu_o$ 和替代假设 $\mu \neq \mu_o$ 为其中的每个样本执行了单样本 t 检验。对于每个检验，我们将假设平均值 μ_o 设置为样本母体的真实平均值。因此，对于给定样本量，10,000 个重复样本中否定原假设的样本将代表模拟 I 类错误概率或检验的显著性水平。由于目标显著性水平为 5%，模拟错误概率约为 0.2%。

模拟结果显示在表 1 和表 2 中，并以图形方式显示在图 1 和图 2 中。

结果和汇总

结果（参见表 1 和图 1）表明，当从对称总体中生成样本时，即使样本量较小，检验的模拟显著性水平也接近目标显著性水平。不过，当样本量较小时，包括从受污染的正态分布中生成的小样本，对称重尾分布的检验结果略显保守。从中还看到，当样本量较小时，异常值将会降低检验的显著性水平。不过，当从对称轻尾母体（Beta (3,3) 和均匀分布）中生成少量样本时，将产生相反的影响。在这些情况下，模拟显著性水平略高。

表 1　对从对称总体中生成的样本执行双侧单样本 t 检验得出的模拟显著性水平。目标显著性水平为 α = 0.05。

Dist.	N(0,1)	t(5)	t(10)	Lpl	CN(.9,3)	CN(.8,3)	B(3,3)	U(0,1)
N	对称和重尾						对称和轻尾	
10	0.050	0.046	0.048	0.044	0.043	0.039	0.057	0.057
15	0.051	0.050	0.049	0.049	0.043	0.043	0.053	0.054
20	0.047	0.051	0.051	0.047	0.043	0.044	0.051	0.052
25	0.050	0.047	0.050	0.046	0.046	0.046	0.048	0.050
30	0.053	0.050	0.048	0.043	0.049	0.046	0.050	0.048
35	0.052	0.047	0.049	0.050	0.047	0.045	0.051	0.054
40	0.046	0.052	0.054	0.048	0.046	0.049	0.044	0.050
50	0.050	0.049	0.051	0.048	0.047	0.051	0.053	0.050
60	0.052	0.049	0.053	0.050	0.051	0.056	0.054	0.052
80	0.049	0.050	0.051	0.047	0.047	0.052	0.049	0.049
100	0.050	0.052	0.049	0.051	0.052	0.054	0.051	0.054

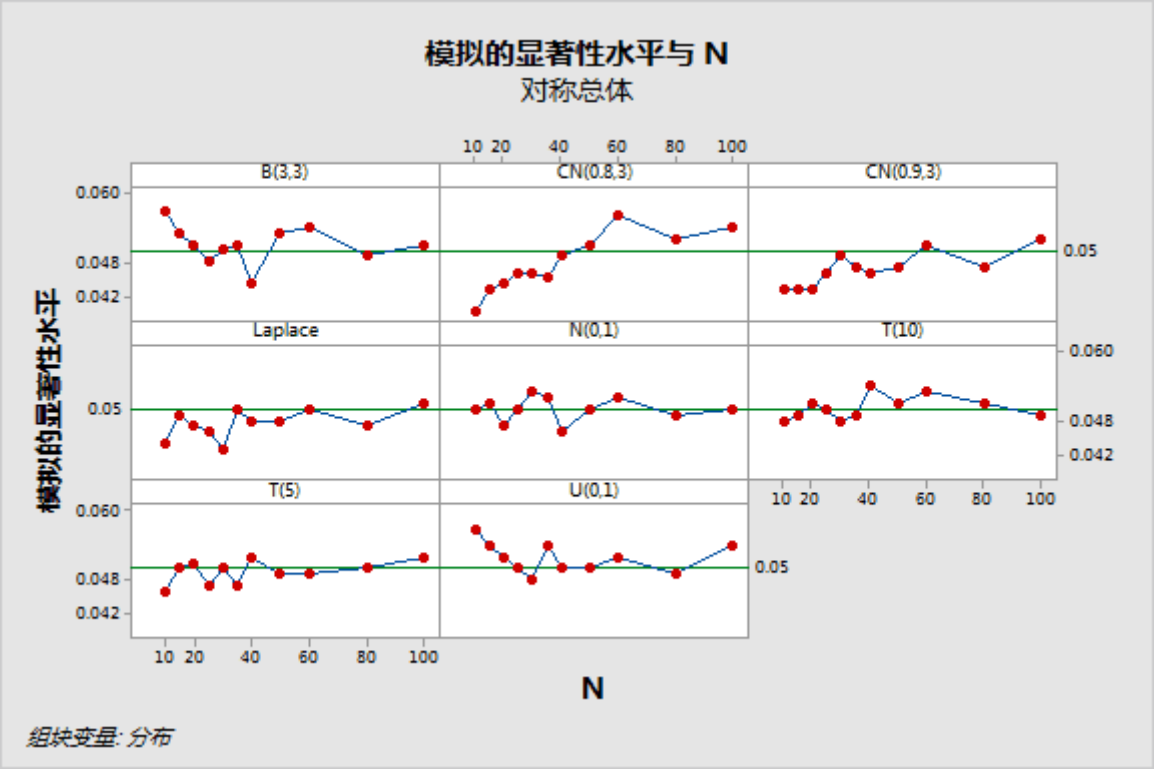

图 1　双侧单样本 t 检验与从对称总体中生成的样本量的模拟显著性水平绘图。目标显著性水平为 α = 0.05。

另一方面，当从偏态分布中生成样本时，检验的效果取决于偏度的严重性。表 2 和图 2 中的结果显示单样本 t 检验对小样本中的偏度敏感。对于重度偏态 总体（指数、Chi (3) 和 Beta(8,1)），则需要较大的样本，模拟显著性水平才能接近目标显著性水平。不过，对于中度偏态总体（Chi(5) 和 Chi(10)），只需提供 20 个样本即可让模拟显著性水平接近目标显著性水平。样本量为 20 时，对于自由度为 5 的卡方分布，模拟显著性水平约为 0.063，对于自由度为 10 的卡方分布，约为 0.056。

表 2　对从偏态总体中生成的样本执行双侧单样本 t 检验得出的模拟显著性水平。目标显著性水平为 $\alpha = 0.05$。

N	期望值	Chi(3)	B(8,1)	Chi(5)	Chi(10)
	总体偏度				
	2.0	1.633	-1.423	1.265	0.894
	模拟显著性水平				
10	0.101	0.089	0.087	0.069	0.060
15	0.088	0.076	0.072	0.068	0.057
20	0.083	0.073	0.069	0.063	0.056
25	0.075	0.068	0.067	0.067	0.056

N	期望值	Chi(3)	B(8,1)	Chi(5)	Chi(10)
	总体偏度				
	2.0	1.633	-1.423	1.265	0.894
	模拟显著性水平				
30	0.069	0.067	0.066	0.058	0.054
35	0.075	0.067	0.062	0.062	0.056
40	0.067	0.067	0.061	0.059	0.056
50	0.064	0.057	0.062	0.057	0.054
60	0.063	0.056	0.061	0.054	0.055
80	0.059	0.058	0.053	0.052	0.052
100	0.060	0.055	0.055	0.047	0.053

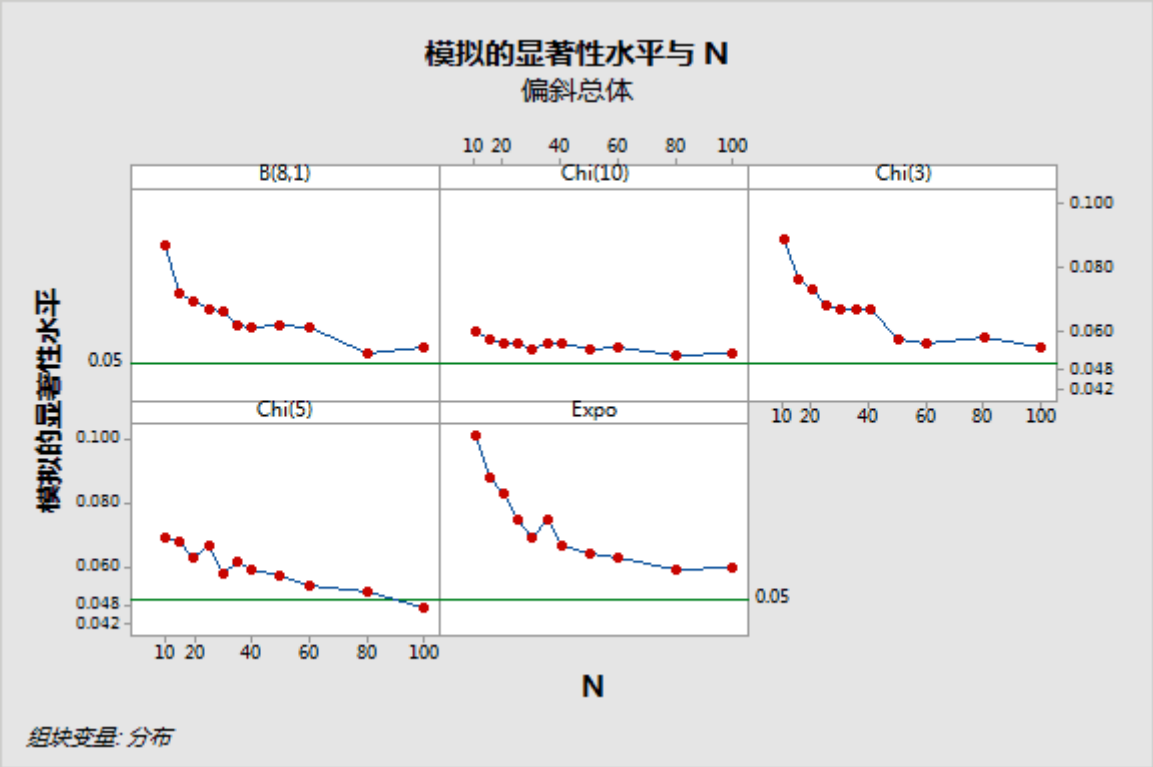

图 2　双侧单样本 t 检验与从偏态总体中生成的样本量的模拟显著性水平绘图。目标显著性水平为 $\alpha = 0.05$。

在本研究中，我们将重点关注假设检验，而不是置信区间。不过，这些结果自然也可延用到置信区间，因为假设检验和置信区间都可以用于确定统计显著性。

附录 B：检验的样本量和功效

我们要研究功效对从中推导出功效的正态假设的敏感度。请注意，如果 β 是 II 类检验错误，则 $1-\beta$ 是检验的功效。因此，我们将确定计划的样本量，以减小 II 类错误概率或大幅提高功效水平。

t 检验的功效众所周知且有据可查。Pearson and Hartley (1954) and Neyman, Iwaszkiewicz, and Kolodziejczyk (1935) 提供功效的图表及表格。

对于样本量为 α 的双侧单样本 t 检验，此功效中有关样本量以及真实平均值 μ 与假设平均值 δ 之间的差值 μ_o 的数学表达式可表示为

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

其中 $F_{d,\lambda}(.)$ 是自由度为 $d = n-1$ 的非中心 t 分布和非中心参数的累积分布函数 (C.D.F)

$$\lambda = \frac{\delta\sqrt{n}}{\sigma}$$

其中，t_d^{α} 代表自由度为 d 的 t 分布的 100 个 α 上百分位点。

对于单侧替代，功效可以表示为

$$\pi(n,\delta) = 1 - F_{n-1,\lambda}(t_{n-1}^{\alpha})$$

用于根据 $\mu > \mu_o$ 检验原假设，可以表示为

$$\pi(n,\delta) = F_{n-1,\lambda}(-t_{n-1}^{\alpha})$$

当根据 $\mu < \mu_o$ 检验原假设时。

这些功效根据以下假设推导出来：数据呈正态分布，检验的显著性水平固定为某个值 α。

模拟研究 B

我们设计了这种模拟，以评估非正态性对单样本 t 检验的理论功效的影响。为了评估非正态性所带来的影响，我们比较了模拟功效水平与使用检验的理论功效计算出的目标功效水平。

我们对从第一项模拟研究中介绍的同一总体中生成的不同大小（$n = 10, 15, 20, 25, 30, 35, 40, 50, 60, 80, 100$）的随机样本执行了差值为 $\alpha = 0.05$ 的双侧 t 检验（参见附录 A）。

对于每一个总体，检验的原假设是 $\mu = \mu_o - \delta$，其替代假设是 $\neq \mu_o - \delta$，其中 μ_o 设置成总体真实平均值和 $\delta = \sigma/2$（σ 是母体的标准差）。这样一来，真实平均值和假设平均值之间的差为 0，因此正确的决定是否定原假设。

对于每一个给定的样本量，10,000 个重复样本从每个分布中提取得到。对于每个给定样本量，10,000 个重复样本中否定原假设的样本代表给定样本量和差值为 δ 的检验的模拟功效水平。请注意，我们之所以选择这个特殊差值，是因为当样本量较小时，它生成的功效值相对较小。

此外，相应的理论功效值（称为目标功效值）根据差值 δ 以及各种样本量计算得到，以与模拟功效值进行比较。

模拟结果显示在表 3 和表 4 中，并以图形方式显示在图 3 和图 4 中。

结果和汇总

结果证实，当样本量足够大时，单样本 t 检验的功效一般对正态假设不敏感。对于从对称总体中生成的样本，表 3 中的结果表明，模拟功效水平接近目标功效水平，即使样本量较小时也如此。图 3 中显示的相应功效曲线几乎没有什么区别。对于从受污染的正态分布中生成的小到中等样本，功效值有点保守。这可能是因为针对那些总体的检验的实际显著性水平比固定的目标显著性水平 α 略高。

表 3　当从对称总体中生成样本时，差值为 $\delta = \sigma/2$，样本量为 $\alpha = 0.05$ 的双侧单样本 t 检验的模拟功效水平。将模拟功效水平与根据正态假设推导出来的理论目标功效水平进行比较。

n	目标功效	N(0,1)	t(5)	t(10)	Lpl	CN(.9,3)	CN(.8,3)	B(3,3)	U(0,1)
		差值为 $\delta = \sigma/2$ 的模拟功效水平（对称总体）							
10	0.293	0.299	0.334	0.311	0.357	0.361	0.385	0.28	0.269
15	0.438	0.438	0.48	0.45	0.491	0.512	0.511	0.423	0.421
20	0.565	0.57	0.603	0.578	0.60	0.629	0.623	0.557	0.548
25	0.67	0.674	0.695	0.68	0.691	0.712	0.70	0.665	0.67
30	0.754	0.756	0.77	0.756	0.767	0.768	0.765	0.754	0.75
35	0.82	0.819	0.827	0.815	0.82	0.819	0.812	0.822	0.818
40	0.869	0.87	0.871	0.868	0.862	0.869	0.868	0.875	0.867
50	0.934	0.933	0.929	0.93	0.929	0.923	0.925	0.932	0.94
60	0.968	0.967	0.963	0.965	0.964	0.955	0.955	0.968	0.971
80	0.993	0.993	0.989	0.992	0.991	0.988	0.989	0.994	0.994
100	0.999	0.998	0.996	0.998	0.999	0.998	0.996	0.999	0.999

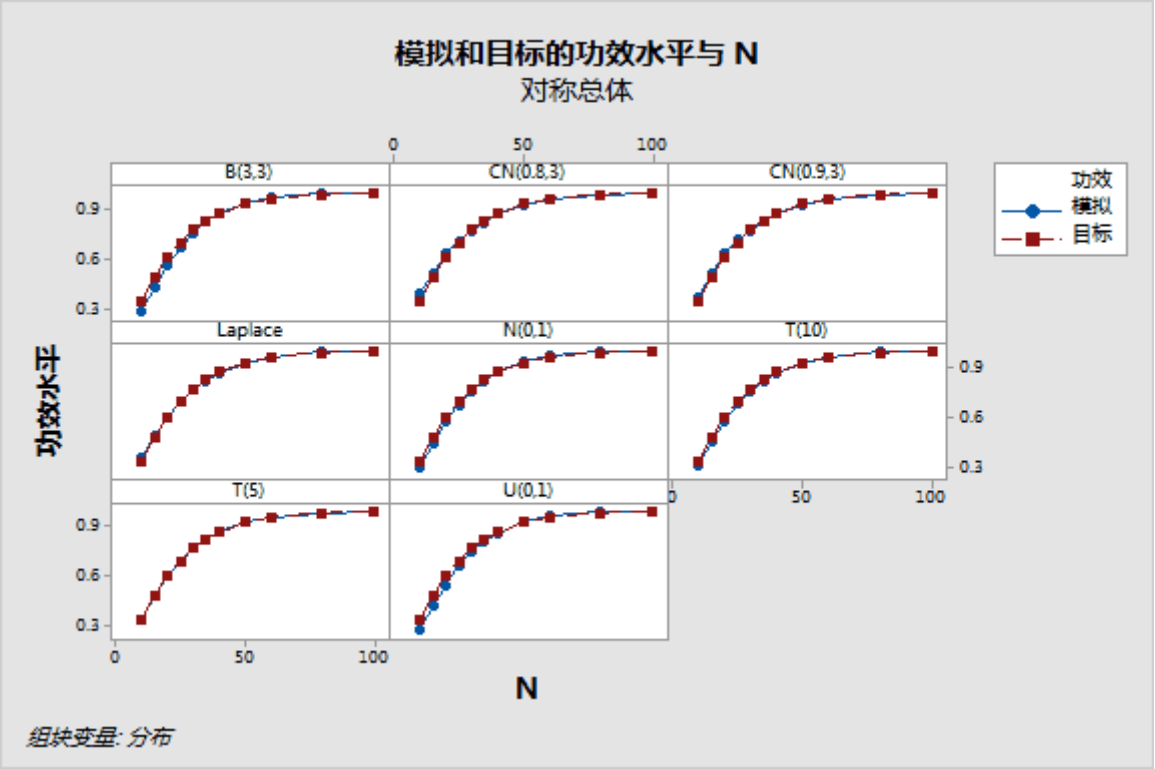

图 3　当从对称总体中生成样本时，差值为 $\alpha = 0.05$ 的双侧单样本 t 检验的模拟功效曲线与理论目标功效曲线比较。根据差值 $\delta = \sigma/2$ 评估功效值。

不过，从偏态总体中生成样本时，小样本的模拟功效值会偏离目标功效值，如表 4 和图 4 中所示。对于中度偏态总体，如自由度分别为 5 和 10 的卡方分布，当样本量至少为 20 时，模拟功效水平才接近目标功效水平。例如，对于 $n = 20$，当卡方 5 和卡方 10 分布的模拟功效水平分别为 0.576 和 0.577 时，目标功效水平为 0.565。对于极度偏态分布，需要较大的样本，才能使模拟功效水平接近目标显著性水平。这可能是因为当样本量较小且母体呈极度偏态分布时，单样本 t 检验无法正确地控制 I 类错误。

表 4　当从偏态总体中生成样本时，差值为 $\delta = \sigma/2$，样本量为 $\alpha = 0.05$ 的双侧单样本 t 检验的模拟功效值。将模拟功效值与根据正态假设推导出的目标功效值进行比较。

N	目标功效	期望值		Chi(3)	B(8,1)	Chi(5)	Chi(10)
			总体偏度				
		2.0		1.633	-1.423	1.265	0.894
			模拟功效水平				
10	0.293	0.206		0.212	0.39	0.225	0.238
15	0.438	0.416		0.413	0.484	0.409	0.407
20	0.565	0.604		0.591	0.566	0.576	0.577
25	0.67	0.763		0.734	0.657	0.709	0.695

N	目标功效	期望值		Chi(3)	B(8,1)	Chi(5)	Chi(10)
			总体偏度				
		2.0		1.633	-1.423	1.265	0.894
			模拟功效水平				
30	0.754	0.859		0.834	0.729	0.808	0.785
35	0.82	0.917		0.895	0.776	0.874	0.835
40	0.869	0.955		0.935	0.823	0.925	0.905
50	0.934	0.987		0.981	0.90	0.973	0.96
60	0.968	0.997		0.994	0.937	0.991	0.985
80	0.993	1.00		0.999	0.98	0.999	0.997
100	0.999	1.00		1.00	0.994	1.00	1.00

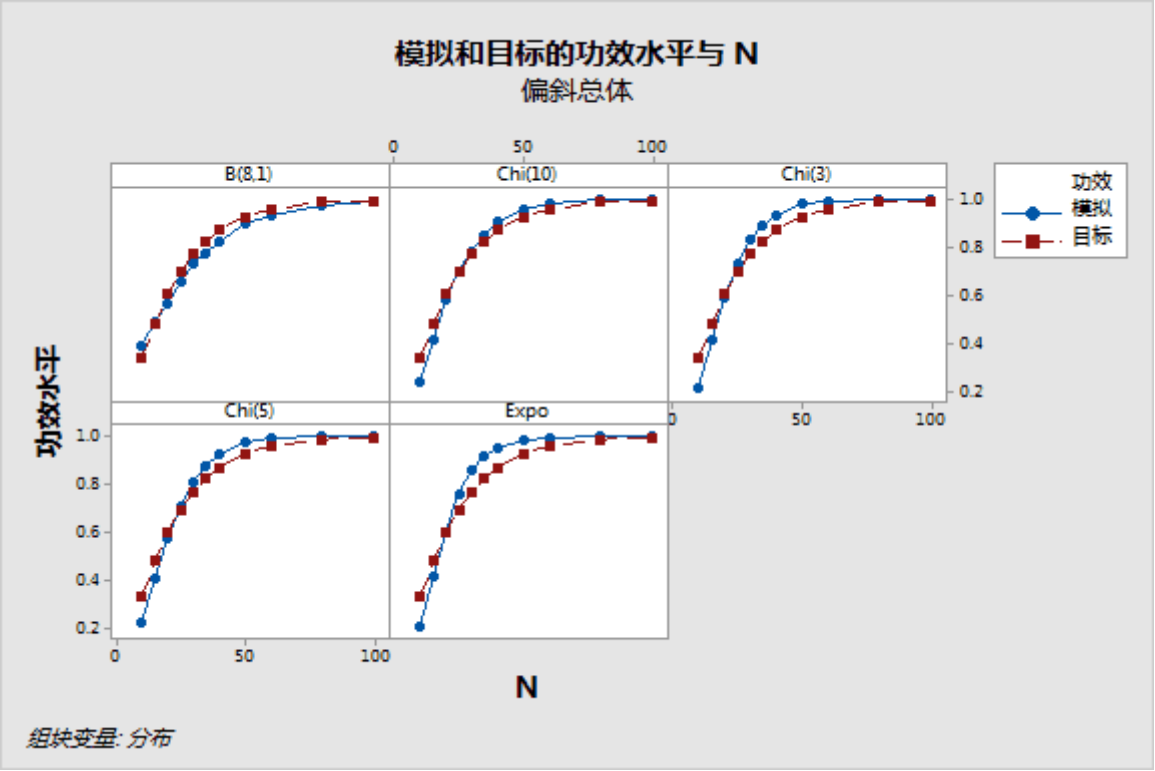

图 4　当从对称总体生成样本时，差值为 $\alpha = 0.05$ 的双侧单样本 t 检验的模拟功效曲线与理论目标功效曲线比较。根据差值 $\delta = \sigma/2$ 评估功效值。

综上所述，对于中度偏态分布，如果样本量至少为 20，那么，无论从哪个母体提取样本，功效都将是可靠的。对于极度偏态总体，需要更大的样本量（约 40），才能使模拟功效接近目标功效。