MINITAB アシスタントホワイトペーパー
本書は、Minitab 統計ソフトウェアのアシスタントで使用される方法およびデータチェックを開発するため、Minitab の統計専門家によって行われた調査に関する一連の文書の 1 つです。

2 サンプル t 検定

概要

2 サンプル t 検定は、2 つの独立したグループが異なるかどうかを比較するために使用できます。この検定は、両方の母集団が正規分布で、等分散性を持つという仮定のもとで導き出されます。正規性の仮定は重要ではありませんが（Pearson 1931、Barlett 1935、Geary 1947）、サンプルサイズが著しく異なる場合、等分散の仮定は重要です（Welch 1937、Horsnell 1953）。

中には、通常の 2 サンプル t 検定手順を実行する前に、等分散を評価するために予備検定を実行する専門家もいます。ただし、これらの分散検定は重要な仮定と制限によって左右されるため、この手法には重大な欠点があります。たとえば、通常の F 検定など、等分散検定の多くは正規性からの逸脱の影響を受けます。Levene/Brown-Forsythe など、正規性の仮定に依存しないその他の検定は、分散間の差を検出する検出力が低すぎます。

B.L. Welch は、分散が必ずしも等しくない場合に、2 つの独立した正規母集団の平均を比較する近似法を開発しました（Welch 1947）。 Welch の修正された t 検定は、分散が等しいという仮定のもとに導き出されるわけではないので、最初に等分散の検定を実行しなくても 2 つの母集団の平均を比較できます。

本書では、Welch の修正された方法による t 検定と通常の 2 サンプル法による t 検定を比較して、どちらの手順が信頼できるかを判断します。また、自動的に実行され、アシスタントレポートカードに表示される以下のデータチェックについて述べ、それらが分析結果にどのように影響するのかについて説明します。

- 正規性
- 異常なデータ
- サンプルサイズ

2 サンプル t 検定方法

通常の 2 サンプル t 検定と Welch の t 検定

データが同じ分散を持つ 2 つの正規母集団から得られたものである場合、通常の 2 サンプル t 検定には、Welch の t 検定と同等またはそれ以上の検出力があります。正規性の仮定は通常の手順では重要でありませんが（Pearson 1931、Barlett 1935、Geary 1947）、確実に有効な結果を得るためには等分散の仮定が重要です。より具体的に言うと、通常の手順では、サンプルサイズが異なる場合、サンプルの大きさに関係なく、等分散の仮定による影響を受けます（Welch 1937、Horsnell 1953）。ただし実際には、等分散の仮定が当てはまることはまれで、その結果、タイプ I 過誤率が高くなる可能性があります。したがって、2 つのサンプルの分散が異なる場合に典型的な 2 サンプル t 検定を使用すると、不正確な結果が出る可能性が高くなります。

Welch の t 検定は、分散が等しいと仮定しないため、あらゆるサンプルサイズの不等分散による影響を受けないので、通常の t 検定の代わりに使用できる実用的な方法です。ただし、Welch の t 検定は近似に基づいており、小さいサンプルサイズでの性能は疑わしい場合があります。アシスタントで使用する検定として、Welch の t 検定と通常の 2 サンプル t 検定のどちらが信頼でき、実用的なのかを判断しようと考えました。

目的

シミュレーションの分析と理論的に導き出すことにより、Welch の t 検定と通常の 2 サンプル t 検定のどちらが信頼できるのかを判断する。より具体的には、次を調べる必要があります。

- データが正規分布に従い、分散が等しい場合、さまざまなサンプルサイズの通常の 2 サンプル t 検定と Welch の t 検定両方のタイプ I およびタイプ II 過誤率。
- 通常の 2 サンプル t 検定が失敗するアンバランス型不等分散計画の Welch の t 検定のタイプ I およびタイプ II 過誤率。

方法

シミュレーションでは、次の 3 つに焦点を当てました。

- 正規性、非正規性、等分散、不等分散、バランス型、アンバランス型計画など、さまざまなモデルを仮定し、通常の 2 サンプル t 検定と Welch の t 検定の検定シミュレーションをした結果を比較しました。詳細は、「付録 A」を参照してください。
- Welch の t 検定の検出力関数を導き出し、通常の 2 サンプル t 検定の検出力関数と比較しました。詳細は、「付録 B」を参照してください。
- Welch の t 検定の理論上の検出力関数に対する非正規性の影響を調査しました。

結果

通常の 2 サンプル t モデルの仮定が有効である場合、Welch の t 検定は小さなアンバランス型計画以外で、通常の 2 サンプル t 検定と同程度またはほぼ同程度に正しく行われます。ただし、通常の 2 サンプル t 検定は、計画が小さくアンバランス型の場合、等分散の仮定の影

響を受けるため、正しく実行されない可能性があります。そのうえ、実際には、2 つの母集団の分散がまったく同じであることを作り出すことは困難です。したがって、Welch の t 検定に対する通常の 2 サンプル検定の理論上の優位性には、実用的な価値がほとんどありません。この理由から、アシスタントでは、2 つの母集団の平均の比較に Welch の t 検定を使用します。シミュレーション結果の詳細は、「付録 A」、「付録 B」、「付録 C」を参照してください。

データチェック

正規性

2 つの独立した母集団の平均を比較するためにアシスタントで使用される Welch の t 検定は、母集団が正規分布に従っているという仮定のもとで導き出されます。また、サンプルが十分に大きければ、Welch の t 検定はデータが正規分布に従っていない場合でも正しく行われます。

目的

Welch の t 検定と典型的な 2 サンプル t 検定のシミュレートした有意水準が、目標有意水準（第一種過誤率）の 0.05 にどれぐらい近いかを判断する。

方法

正規母集団、歪んだ母集団、および混合正規母集団（等分散および不等分散）から生成されたサンプルペア 10,000 個で Welch の t 検定と通常の 2 サンプル t 検定のシミュレーションを行いました。サンプルは、さまざまなサイズのものを使用しました。正規母集団は、比較対照母集団の役目を果たします。各条件でシミュレートした有意水準を計算し、目標（名目）有意水準の 0.05 と比較しました。検定が正しく行われれば、シミュレートした有意水準は 0.05 に近くなります。

結果

中規模または大きなサンプルの場合、Welch の t 検定では正規データおよび非正規データのタイプ I 過誤率が維持されました。両方のサンプルサイズが少なくとも 15 のとき、シミュレートされた有意水準は目標有意水準に近くなります。詳細は、「付録 A」を参照してください。

検定が比較的小さなサンプルで正しく行われるため、アシスタントではデータの正規性は検定されません。代わりに、サンプルのサイズを調べ、レポートカードに次のステータスインジケータが表示されます。

ステータス	状態
	両方のサンプルサイズが少なくとも 15 です。正規性に問題はありません。
	少なくとも 1 つのサンプルサイズが 15 未満です。正規性に問題がある可能性があります。

異常なデータ

異常なデータとは、極端に大きいまたは小さいデータ値を指し、外れ値とも呼ばれています。異常なデータは、分析の結果に強い影響を与える可能性があります。サンプルサイズが小さい場合は、統計的に重要な結果を見つける機会に影響を与える場合が有ります 異常なデータは、データ収集での問題や工程の異常な動作を示している可能性があります。したがって、これらのデータ点は調査する価値があり、可能な場合には修正する必要があります。

目的

分析の結果に影響する可能性がある、全体のサンプルに比べて非常に大きい、または非常に小さいデータ値をチェックする方法を開発する必要がありました。

方法

箱ひげ図の外れ値を特定するために、Hoaglin, Iglewicz, and Tukey (1986) によって提唱された方法に基づいて、異常データをチェックする方法を開発しました。

結果

アシスタントでは、分布の下位四分位数を下回る幅または上位四分位数を上回る幅が四分位間範囲の 1.5 倍より大きいデータ点は異常と識別されます。下位および上位四分位数とは、データの 25 番目および 75 番目の百分位数を指します。四分位間範囲とは、2 つの四分位数間の差を指します。この方法は、特定の各外れ値を検出することが可能なため、複数の外れ値がある場合でも有効に機能します。

外れ値は、サンプルサイズが非常に小さい場合にのみ、検出力関数に影響を与える傾向があります。一般に、外れ値がある場合、観測された検出力は目標とする理論上の検出力より少し高くなる傾向があります。このパターンは「付録 C」の図 10 で確認できます。最小サンプルサイズが 15 に達するまでは、シミュレートした検出力曲線と理論検出力曲線は離れており近接しません。

異常なデータを調べる場合、2 サンプル t 検定のアシスタントレポートカードには次のステータスインジケータが表示されます。

ステータス	状態
	異常なデータ点はありません。
	少なくとも 1 つのデータ点が異常です。検定結果に影響する可能性があります。

サンプルサイズ

一般的に、仮説検定は、「差がない」という帰無仮説を棄却する証拠を集めるために実行されます。サンプルが小さすぎると、実際に平均に差があるときに、その差を検出ための検定の検出力が十分でない場合あり、その結果、タイプ II 過誤が生じます。したがって、サンプルサイズが実質的に重要な差を検出するのに十分な大きさであることを確認することが重要です。

目的

現在のデータから帰無仮説に反する十分な証拠が得られない場合、検定のサンプルサイズが、高い確率で対象となる実質的な差を検出するのに十分な大きさかどうかを判断する必要があります。サンプルサイズ計画の目的は、サンプルが、重要な差を高い確率で検出するために十分な大きさであることを確認することですが、サンプルサイズが大きすぎて無意味な差が高い確率で統計的に有意になってはなりません。

方法

検出力とサンプルサイズ分析は、統計分析を実行する際使用される特定の検定の理論検出力関数に基づきます。Welch の t 検定の場合、この検出力関数はサンプルサイズ、2 つの母集団の平均間の差、および 2 つの母集団の実際の分散によって異なります。詳細は、「付録 B」を参照してください。

結果

データから帰無仮説に反する十分な証拠が得られない場合、アシスタントでは与えられたサンプルサイズで 80%および 90%の確率で検出できる実質的な差が計算されます。さらに、ユーザーが対象の実質的な差を指定すると、80%および 90%の確率で差を検出できるサンプルサイズが計算されます。

結果はユーザーの特定のサンプルによって異なるため、一般的な結果はなにも報告されません。Welch 検定の検出力関数についての詳細は、「付録 B」と「付録 C」を参照してください。

検出力とサンプルサイズを調べる場合、2 サンプル t 検定のアシスタントレポートカードには次のステータスインジケータが表示されます。

ステータス	状態
	検定で平均間の差が検出されます。検出力に問題はありません。 または 検出力は十分です。検定で平均間の差は検出されませんでしたが、サンプルは少なくとも 90%の確率で差を検出するのに十分な大きさです。
	検出力は十分である可能性があります。検定で平均間の差は検出されませんでしたが、サンプルは少なくとも 80～90%の確率で差を検出するのに十分な大きさです。90%の検出力を達成するのに必要なサンプルサイズが報告されます。

ステータス	状態
	検出力は十分でない可能性があります。検定で平均間の差は検出されませんでしたが、サンプルは少なくとも 60～80%の確率で差を検出するのに十分な大きさです。80%および 90%の検出力を達成するのに必要なサンプルサイズが報告されます。
	検出力は十分ではありません。検定で平均間の差は検出されませんでした。サンプルは少なくとも 60%の確率で差を検出するのに十分な大きさではありません。80%および 90%の検出力を達成するのに必要なサンプルサイズが報告されます。
	検定で平均値間の差が検出されませんでした。ユーザーが検出すべき実際の平均の差を指定しなかったので、サンプルサイズ、標準偏差および α に基づいて 80%および 90%の確率で検出可能な差がレポートに示されます。

参考文献

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. (1958). 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.

付録 A: 通常の 2 サンプル t 検定と Welch の t 検定に対する非正規性と異質性の影響

異なるモデル仮定のもとで、通常の 2 サンプル t 検定と Welch の t 検定を比較するシミュレーションの分析を行いました。

シミュレーションの調査 A

次の 3 つに分けて調査を行いました。

- 調査の第 1 部では、正規性の仮定が当てはまる場合に、等分散の仮定による通常の 2 サンプル t 検定と Welch の t 検定への影響を調査しました。2 つのサンプルは 2 つの独立する正規母集団から生成されました。1 つ目のサンプル（基本サンプル）は平均 0、標準偏差$\sigma_1 = 2$、$N(0,2)$の正規母集団から抽出されました。2 つ目のサンプルも平均 0 の正規母集団から抽出されましたが、比$\rho = \sigma_2/\sigma_1$が 0.5、1.0、1.5 となるように選ばれた標準偏差σ_2を使用しました。つまり、2 つ目のサンプルはそれぞれ母集団$N(0,1)$、$N(0,2)$、$N(0,3)$、$N(0,4)$から抽出されました。さらに、各ケースの基本サンプルサイズは与えられた各n_1で$n_1 = 5, 10, 15, 20$に固定されました。2 つ目のサンプルサイズn_2は、サンプルサイズの比$r = n_2/n_1$が 0.5、1、1.5、2.0 にほぼ等しくなるように選択されました。

  これらの 2 サンプル計画のそれぞれに、各母集団から独立したサンプルペア 10,000 個を生成しました。次に、平均間に差がないという帰無仮説を検定するために、各 10,000 個のサンプルペアに通常の 2 サンプル t 検定と Welch の t 検定を実行しました。平均間の真の差はゼロであるため、反復数 10,000 のうち帰無仮説が棄却される割合は、検定のシミュレートした有意水準を表します。各検定の目標有意水準は$\alpha = 0.05$であるため、各検定と各実験に関連付けられたシミュレーション誤差は約 0.2% です。

- 第 2 部では、2 つの検定のシミュレートした有意水準に対する非正規性、特に歪みによる影響を調査しました。このシミュレーションは、以下の点を除き、前のシミュレーションと同様に設定されました。基本サンプルは自由度 2 のカイ二乗分布（Chi(2)）から抽出されました。2 つ目のサンプルは$\rho = \sigma_2/\sigma_1$が 0.5、1.0、1.5、2 の値をとるように他のカイ二乗分布から抽出されました。仮説の平均間の差は、親母集団の平均間の真の差に設定されました。

- 第 3 部では、2 つの t 検定の性能に対する外れ値の影響を調査しました。そのため、2 つのサンプルは混合正規分布から抽出されました。混合正規分布$\mathrm{CN}(\mathrm{p}, \sigma)$は、$\mathrm{N}(0,1)$母集団と正規$\mathrm{N}(0, \sigma)$母集団という 2 つの正規母集団の混合です。混合正規分布を次のように定義します。

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

ここで、pは混合パラメータ、$1-p$は混入の比率（外れ値の比率）です。Xが$CN(p,\sigma)$として分布される場合、平均が$\mu_X = 0$、標準偏差が$\sigma_X = \sqrt{p+(1-p)\sigma^2}$であることが簡単に示されます。

基本サンプルは$CN(.8,4)$から、2つ目のサンプルは混合正規$CN(.8,\sigma)$から抽出されました。パラメータσは、第I部、第II部と同様に2つの（混合）分布$\rho = \sigma_2/\sigma_1$の標準偏差の比率が0.5、1.0、1.5、2に等しくなるように選択されました。$\sigma_1 = \sqrt{.8+(1-.8)*16} = 2.0$なので、$\sigma = 1,4,6.40,8.72$がそれぞれ選択されます。つまり、2つ目のサンプルは$CN(.8,1)$、$CN(.8,4)$、$CN(.8,6.4)$、$CN(.8,8.72)$から抽出されました。次に、第I部で説明したようにシミュレーションを行いました。

分析の結果を表1にまとめ、図1、図2、図3に示します。

結果と要約

一般に、シミュレーションの結果は、正規性および等分散の仮定のもとで、サンプルサイズが小さいとき、通常の2サンプルt検定の有意水準は目標水準に近くなるという理論上の結果を裏付けています。図1の第2列のプロットは、2つの正規母集団の分散が等しい計画のシミュレートした有意水準を示しています。通常の2サンプルt検定に基づくシミュレートした有意水準の曲線は、目標水準の線と区別できません。

下の表は、通常の2サンプルt検定とWelchのt検定両方の両側検定についてシミュレートした有意水準を示しており、正規母集団、歪んだ母集団（カイ二乗）、混合正規母集団から生成された一対のサンプルに基づく$\alpha = 0.05$がそれぞれ使用されています。サンプルのペアは同じ分布族からのものですが、それぞれの親母集団の分散は必ずしも等しくありません。

表1　n = 5の両側検定（$\alpha = 0.05$を使用した通常の2サンプルt検定とWelchのt検定）のシミュレートした有意水準

			基本母集団：N(0,2) 第2母集団：N(0, σ_2)				基本母集団：Chi(2) 第2母集団：カイ二乗				基本母集団：CN(.8,4) 第2母集団：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 の両側検定（$\alpha = 0.05$を使用した通常の 2 サンプル t 検定と Welch の t 検定）のシミュレートした有意水準

			基本母集団：N(0,2) 第 2 母集団：N(0, σ_2)				基本母集団：Chi(2) 第 2 母集団：カイ二乗				基本母集団：CN(.8,4) 第 2 母集団：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 の両側検定（$\alpha = 0.05$を使用した古典的 2 サンプル t 検定と Welch の t 検定）のシミュレートした有意水準

			基本母集団：N(0,2) 第 2 母集団：N(0, σ_2)				基本母集団：Chi(2) 第 2 母集団：カイ二乗				基本母集団：CN(.8,4) 第 2 母集団：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

			基本母集団：N(0,2) 第 2 母集団：N(0, σ_2)				基本母集団：Chi(2) 第 2 母集団：カイ二乗				基本母集団：CN(.8,4) 第 2 母集団：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	$n_1 = 15$				$n_1 = 15$				$n_1 = 15$			
30	2.0	2T	.111	.050	.028	.018	.123	.049	.027	.020	.100	.046	.025	.016
		Welch	.049	.051	.051	.053	.074	.056	.045	.047	.039	.044	.042	.040

表 4　n = 20 の両側検定（$\alpha = 0.05$を使用した通常の 2 サンプル t 検定と Welch の t 検定）のシミュレートした有意水準

			基本母集団：N(0,2) 第 2 母集団：N(0, σ_2)				基本母集団：Chi(2) 第 2 母集団：カイ二乗				基本母集団：CN(.8,4) 第 2 母集団：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

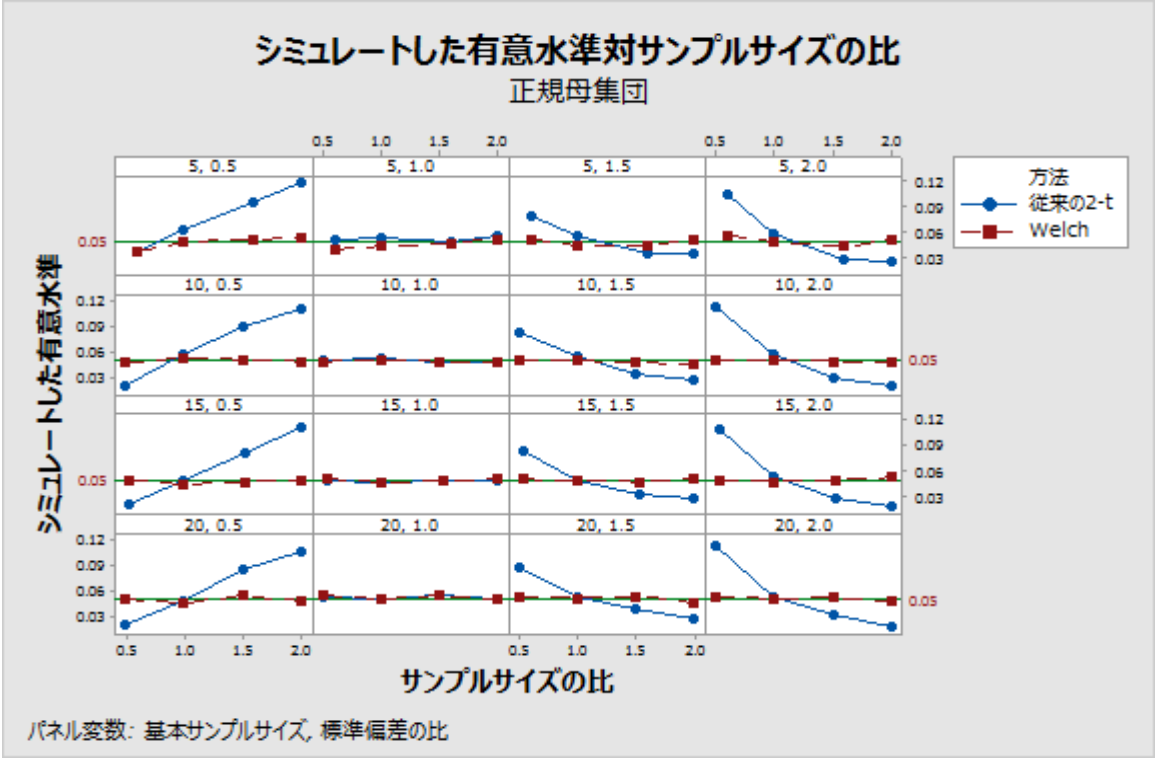

図 1　等分散または不等分散がサンプルサイズ比に対しプロットされた 2 つの正規母集団から生成された一対のサンプルに基づく両側検定（$\alpha = 0.05$を使用した通常の 2 サンプル t 検定と Welch の t 検定）のシミュレートした有意水準。

シミュレーション結果は、比較的小さなサンプルの場合、通常の 2 サンプル t 検定は非正規性に対して頑健ですが、2 サンプル計画がほぼバランス型でなければ、等分散の仮定による影響を受けることを示しています。これは、図 1、図 2、図 3 のグラフで示されています。通常の 2 サンプル t 検定に基づくシミュレートした有意水準の曲線は、分散が非常に異なる場合でも、サンプルサイズ比が 1.0 の点で目標水準の線と交差します。3 つすべての分布族（正規、カイ二乗、および混合正規母集団）で、サンプルサイズが異なる場合、通常の 2 サンプル t 検定のシミュレートした有意水準は、分散が等しいとき目標水準に近くなります。これは、図 1、図 2、図 3 それぞれの第 2 列のプロットに示されています。

計画がアンバランス型で分散が等しくない場合、通常の t 検定は望ましくない性能を示しています。分散間の小さな格差でも問題となります。不等分散のアンバランス型計画では、シミュレートした有意水準がデータの正規性によって向上することはありません。実際、親母集団に関係なく、サンプルサイズが増加するにつれ、シミュレートした有意水準は目標水準から離れていきます。分散がより大きい母集団からより大きなサンプルが抽出されると、シミュレートした有意水準は目標水準より小さくなります。分散がより小さい母集団からより大きなサンプルが抽出されると、シミュレートした有意水準は目標水準より大きくなります。Arnold（1990、372 ページ）は、不等分散仮定のもとの通常の 2 サンプル t 検定統計量の漸近分布を調査したときに、同様のことを述べています。

一方、Welch の 2 サンプル t 検定は、図 1、図 2、図 3 に示すように、等分散仮定からの逸脱に影響されません。Welch の t 検定は等分散の仮定のもとに導出されるわけではないので、これは当然のことです。Welch の t 検定が導き出される正規性の仮定は、2 つのサンプルの最小サイズが非常に小さい場合にのみ重要となるようです。ただし、より大きいサンプルでは、検定は正規性の仮定からの逸脱による影響を受けなくなります。これは図 2 と図 3 に示

されています。2 つのサンプルの最小サイズが 15 のとき、シミュレートした有意水準は目標水準に一貫して近いままでいます。両方のサンプルが、自由度 2 のカイ二乗分布から生成され、サンプルサイズが共に 15 のとき、シミュレートした有意水準は 0.042 になります（図 3 を参照）。

また、2 つのサンプルの最小サイズが十分に大きい場合、外れ値は Welch の t 検定の性能に影響するようには見えません。表 3 と図 3 は、2 つのサンプルの最小サイズが 15 以上のとき、シミュレートした有意水準は目標水準に近くなることを示しています（標準偏差の比が 0.5、1.0、1.5、2.0 のとき、シミュレートした有意水準はそれぞれ 0.045、0.045、0.041、0.037 です）。

これらの結果は、ほとんどの実用目的において、シミュレートした有意水準またはタイプ I 過誤率という点では、Welch の 2 サンプル t 検定は通常の 2 サンプル t 検定よりも優れていることを示しています。

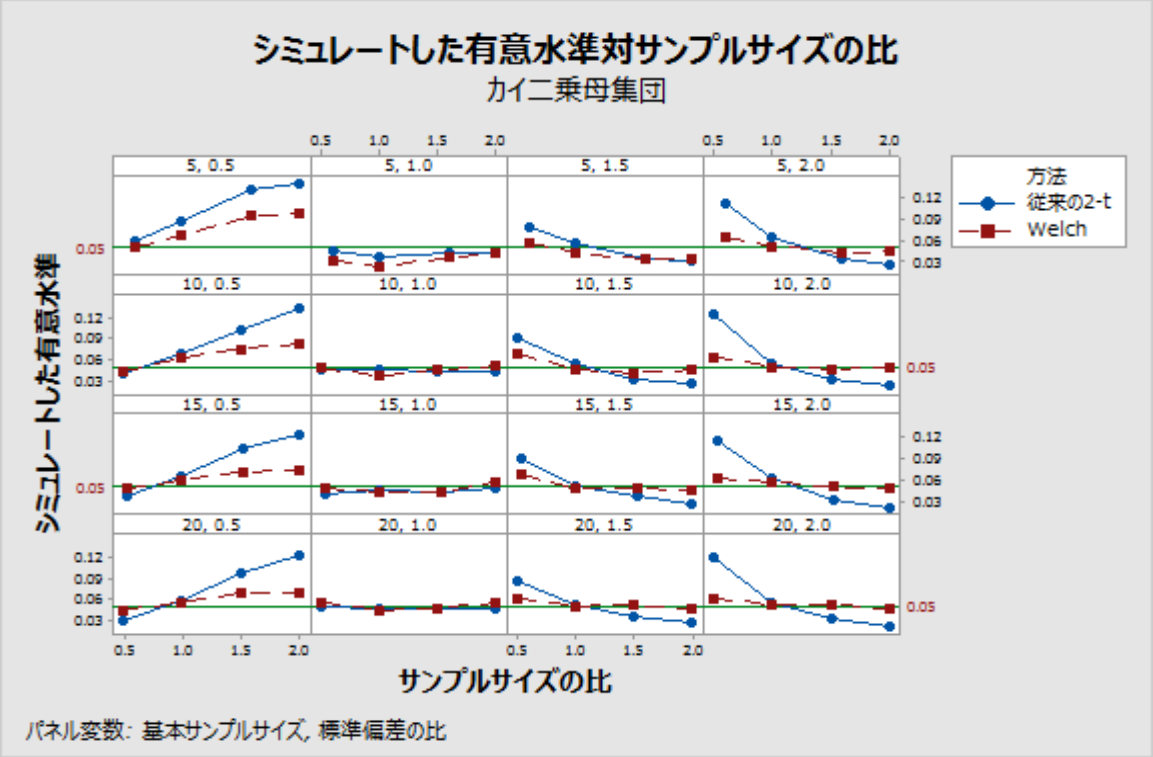

図 2　等分散または不等分散がサンプルサイズ比に対しプロットされた 2 つの正規母集団から生成されたサンプルペアに基づく両側検定（古典的 2 サンプル t 検定と Welch の t 検定）のシミュレートした有意水準。

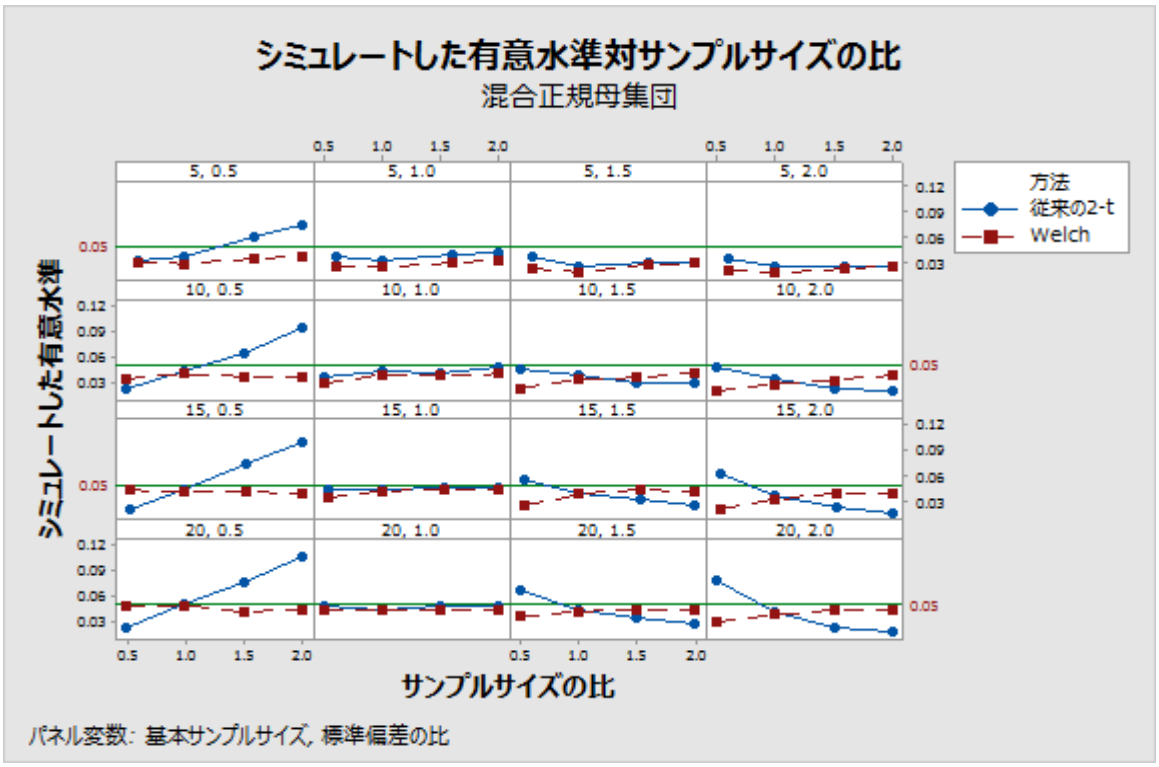

図 3　等分散または不等分散がサンプルサイズ比に対しプロットされた 2 つの正規母集団から生成されたサンプルペアに基づく両側検定（古典的 2 サンプル t 検定と Welch の t 検定）のシミュレートした有意水準。

付録 B：2 つの検定の検出力関数の比較

Welch の t 検定の検出力関数が、通常の 2 サンプル t 検定の検出力関数に等しい、またはほぼ等しくなる条件を決定する必要がありました。

一般に、t 検定（1 サンプルまたは 2 サンプル）の検出力関数はよく知られており、多くの刊行物で言及されています（Pearson and Hartley 1952、Neyman et al 1935、Srivastava 1958）。次の定理は、2 サンプル計画の 3 つの異なる対立仮説それぞれの検出力関数について述べています。

定理 B1

正規性と等分散を仮定したとき、名目サイズ α を持つ両側 2 サンプル t 検定の検出力関数は、サンプルサイズおよび差 $\delta = \mu_1 - \mu_2$ の関数として、次のように表すことができます。

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

ここで、$F_{d_C,\lambda}(.)$ は自由度 $d_C = n_1 + n_2 - 2$ および次の非心パラメータを使用した非心 t 分布の累積分布関数（C.D.F）です。

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

さらに、対立仮説 $\mu_1 > \mu_2$ に関連付けられた検出力関数は次のように与えられます。

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

一方、対立仮説 $\mu_1 < \mu_2$ に対して検定する場合、検出力は次のように表されます。

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

上の定理の結果はよく知られていますが、この文献では Welch の修正された t 検定に基づく検定の検出力については特に言及されていません。近似は、一元配置の分散分析モデルに対して与えられた近似の検出力から推定できます（Kulinskaya et al 2003）。残念ながら、この検出力関数は両側検定の対立仮説にのみ適用されます。ただし、2 サンプル計画は、3 つの対立仮説のそれぞれについて Welch の t 検定の（正確な）検出力関数を得るために異なる手法を採択できる特殊なケースです。これらの関数は、次の定理で与えられます。

定理 B2

母集団が正規分布に従う（ただし分散は必ずしも同じではない）と仮定したとき、名目サイズ α を持つ両側の Welch の t 検定の検出力関数は、サンプルサイズおよび差 $\delta = \mu_1 - \mu_2$ の関数として、次のように表すことができます。

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

ここで、$G_{d,\lambda}(.)$ は次のように与えられる自由度 d_W

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

および次の非心パラメータを使用した非心 t 分布の累積分布関数（C.D.F）です。

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

片側の対立仮説の場合、

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

および

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

として、対立仮説$\mu_1 > \mu_2$に対する帰無仮説、および対立仮説$\mu_1 < \mu_2$に対する帰無仮説を検定するために、検出力関数がそれぞれ与えられます。

結果の証明は、「付録 D」を参照してください。

これらの 2 つの検出力関数を比較する前に、通常の 2 サンプル t 検定は母集団の分散が等しいという追加の仮定のもとに導き出されるため、この 2 番目の仮定が Welch の t 検定に当てはまる場合、2 つの検定の理論上の検出力関数を比較する必要があります。

理論的には、正規性および等分散の仮定のもとでは、次のようになることがわかっています。

すべてのn_1, n_2, δで$\pi(n_1, n_2, \delta) \geq \pi_W(n_1, n_2, \delta)$

次の結果は、2 つの関数が（ほぼ）等しくなる条件を示しています。

定理 B3

正規性および等分散の仮定のもとでは、次のようになります。

1. $n_1 \sim n_2$の場合、各差δについて$\pi(n_1, n_2, \delta) \sim \pi_W(n_1, n_2, \delta)$となります。特に、$n_1 = n_2$の場合、各差$\delta$について$\pi(n_1, n_2, \delta) = \pi_W(n_1, n_2, \delta)$となり、Welch の t 検定は通常の 2 サンプル t 検定と同じくらい強力です。
2. n_1とn_2が小さく、$n_1 \neq n_2$の場合、Welch の t 検定は通常の 2 サンプル t 検定より劣ります。ただし、n_1とn_2が大きい場合、（サンプルサイズ間の差に関係なく）$\pi(n_1, n_2, \delta) \sim \pi_W(n_1, n_2, \delta)$となります。

結果の証明は、「付録 E」を参照してください。

分散が等しいと仮定した場合、2 つの検定の検出力関数に関連付けられた非心パラメータは同一になります。検出力関数の差は、それぞれの自由度の差にのみ起因します。理論上、上記の仮定のもとでは、通常の t 検定は UMP（一様最強力）であることがわかっており、したがって自由度がより高くなります。ただし、上記の結果の要点は、計画がバランス型またはほぼバランス型である場合、検出力関数は同一、またはほぼ同一であるということです。通常の t 検定が、Welch の t 検定より明らかに強力である唯一のケースは、計画が著しくアンバランス型で、サンプルが小さい場合です。古典的 t 検定が、Welch の t 検定より明らかに強力である唯一のケースは、計画が著しくアンバランス型で、サンプルが小さい場合です。残念ながら、「付録 A」に示すように、これは等分散の仮定によって通常の 2 サンプル t 検

定が特に影響を受けるケースでもあります。その結果、Welch の t 検定の検出力は、実用目的にはより信頼できる関数ということになります。

2 つの正規母集団が同じ標準偏差 3 である次の例で、定理 B3 の結果を示します。検出力は、次の 4 つのシナリオで定理 B1 と定理 B2 の（両側）検出力関数に基づいて計算されます。

1. どちらのサンプルも小さいが、同じサンプルサイズである（$n_1 = n_2 = 10$）。
2. どちらのサンプルも小さいが、一方のサンプルは他方のサンプルより 2 倍大きい（$n_1 = 10, n_2 = 20$）
3. 一方のサンプルは小さく、他方のサンプルは中規模である。中規模のサンプルは小さいサンプルより 4 倍大きい（$n_1 = 10, n_2 = 40$）。
4. 一方のサンプルは中規模で、他方のサンプルは大きい。大きいサンプルは中規模のサンプルより 4 倍大きい（$n_1 = 50, n_2 = 200$）。

両方の検定で$\alpha = 0.05$であると仮定して、差$\delta = 0.0, 0.5, 1.0, 1.5, 2.0, \ldots 5.0$で各シナリオの検出力が評価されます。表 5 に結果を示し、図 4 に関数をプロットします。

表 5　通常の 2 サンプル t 検定と Welch の t 検定の両側検定における理論上の検出力関数の比較（$\alpha = 0.05$）。サンプルサイズn_1とn_2が固定され、検出力関数が 0.0～0.5 の差δで評価されます。

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.

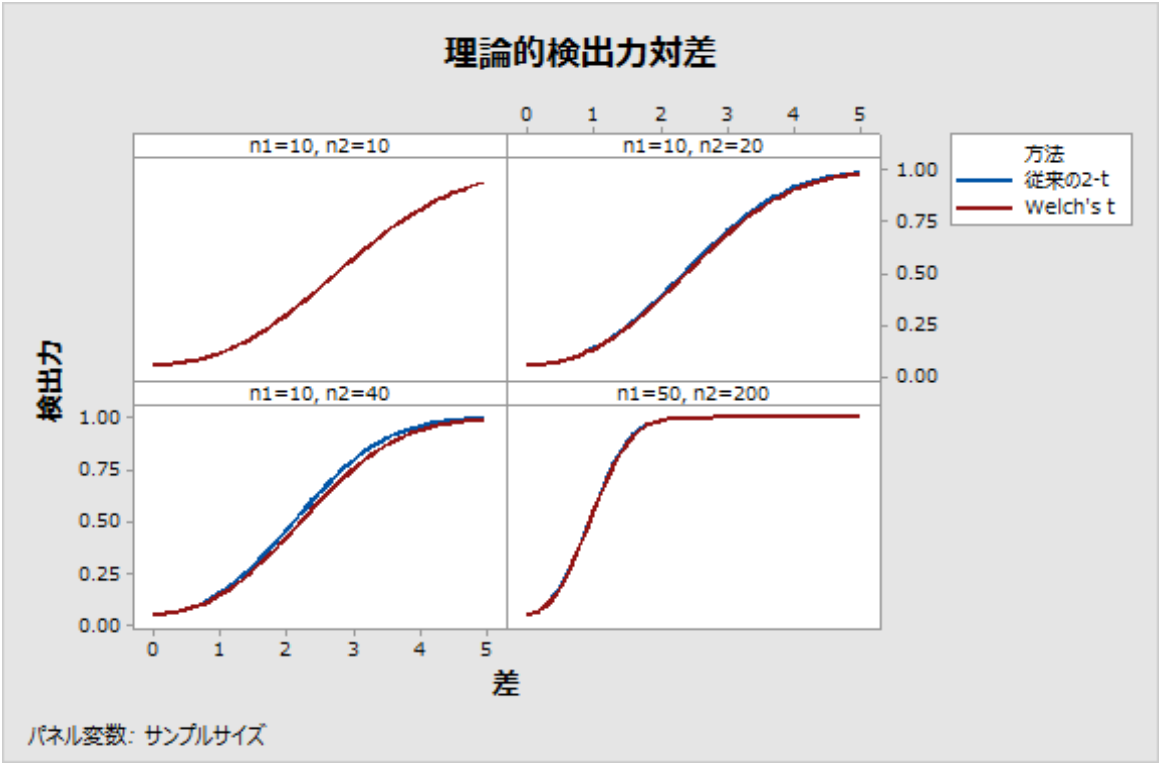

図 4　検出される平均間の差δに対し、両側の通常の 2 サンプル t 検定および両側の Welch の t 検定の理論上の検出力を示すプロット。両方の検定で$\alpha = 0.05$が使用されています。母集団は正規母集団で標準偏差は 3 で等しいと仮定します。

シミュレーションの調査 B

このシミュレーションの調査の目的は、分散が等しくないと仮定するバランス型計画で、通常の 2 サンプル t 検定に関連付けられた検出力水準と、Welch の 2 サンプル t 検定に関連付けられた検出力水準を比較することです。これらの分析での実験は、「付録 A」で説明した実験と類似しています。

最初の実験グループでは、分散が等しくない正規母集団から等しいサイズのサンプルペアを生成しました。基本母集団は$N(0,2)$に固定され、2 番目の正規母集団は標準偏差$\rho = \sigma_2/\sigma_1$が 0.5、1.5、2 になるように選択されました。同様に、2 番目のグループでは、分散が等しくないカイ二乗分布から 2 つのサンプルが抽出されました（基本母集団は Chi(2)）。最後の実験のグループでは、「付録 A」で前述したとおり、混合正規分布（基本母集団 CN(.8,4)）から一対のサンプルが生成されました。

各実験グループで、サンプルサイズ$n = n_1 = n_2 = 5, 10, 15, 20, 25, 30$の各検定に関連付けられた、シミュレートした検出力水準を（所与の検出可能な差δで）計算しました。各実験で、シミュレートした検出力水準は、帰無仮説が正しくないときに棄却される場合の比率として計算されました。すべての実験で、平均間の差は基本母集団（2 つあるうち 1 番目のサンプル）の標準の 1 単位で指定されました。より具体的に言うと、この分析の 3 つの分布族すべてで比較的小さいため、$\delta = 1.0 \times \sigma_1 = 2.0$に固定しました。表 2.2、図 2.2a、図 2.2b、図 2.2c にシミュレーション結果を示します。

結果と要約

表 6 と図 4 の結果は、定理 2.3 で示すように、等分散の仮定のもとでは、バランス型計画の理論上の検出力関数は同一になることを示しています。さらに、サンプルサイズが比較的小さいけれども、ほぼ同じサイズの場合は、2 つの関数の検出力はほとんど同じになります。サンプルサイズが小さく、一方のサンプルが他方の約 4 倍の大きさになって初めて（たとえば、$n_1 = 10, n_2 = 40$）、検出力関数間に明らかな差が見られるようになります。その場合でも、通常の 2 サンプル t 検定に基づく理論上の検出力は、Welch の t 検定に基づく検出力より若干高くなるだけです。最後に、計画が著しくアンバランス型であっても、サンプルが（比較的）大きい場合は、定理 B3 で述べたとおり、2 つの検出力関数は基本的に同一です。

さらに、分散が等しくないバランス型計画では、2 つの検定の検出力は実質的に同一になります。ただし、非常に小さいサンプル（$n < 10$）では、通常の 2 サンプル t 検定のほうがやや優れています。

表 6　バランス型不等分散計画での通常の 2 サンプル t 検定と Welch の t 検定のシミュレートした検出力水準の比較

		基本母集団：N(0,2)			基本母集団：Chi(2)			基本母集団：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

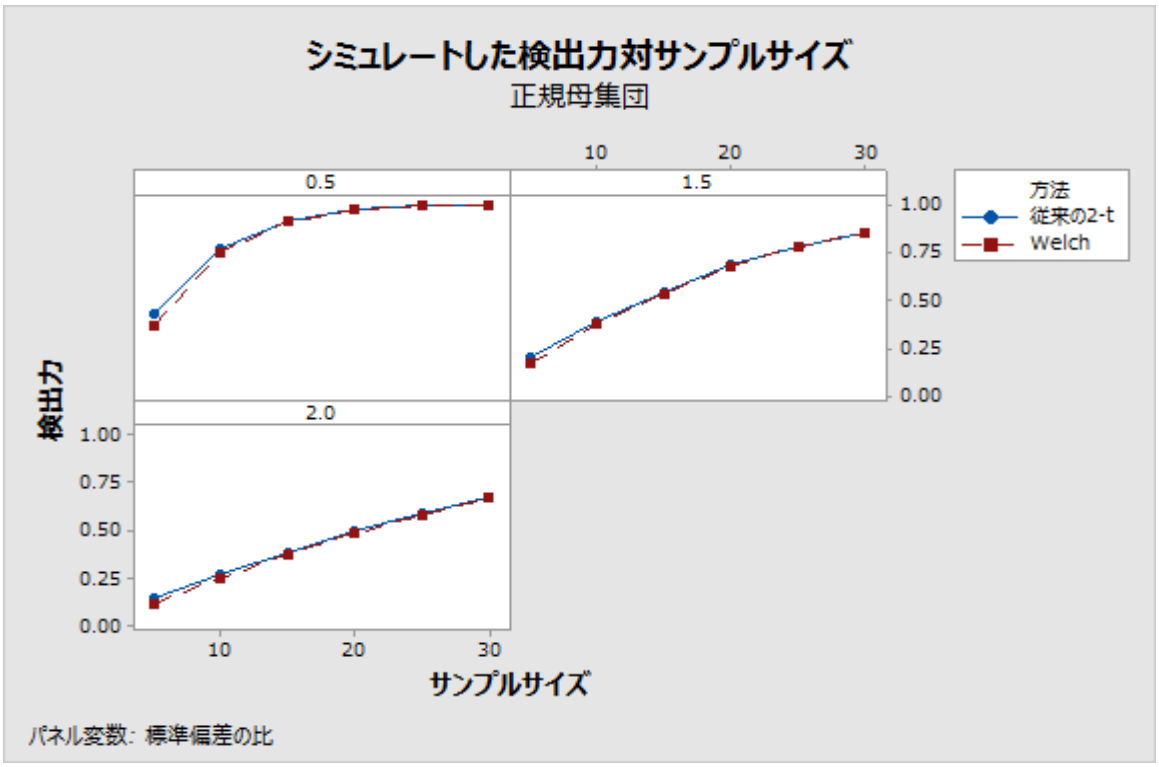

図 5　バランス型不等分散計画での通常の 2 サンプル t 検定と Welch の 2 サンプル t 検定のシミュレートした検出力水準の比較。サンプルは、標準偏差の比が 0.5、1.5、2.0 になるように、分散が等しくない正規母集団から抽出されました。

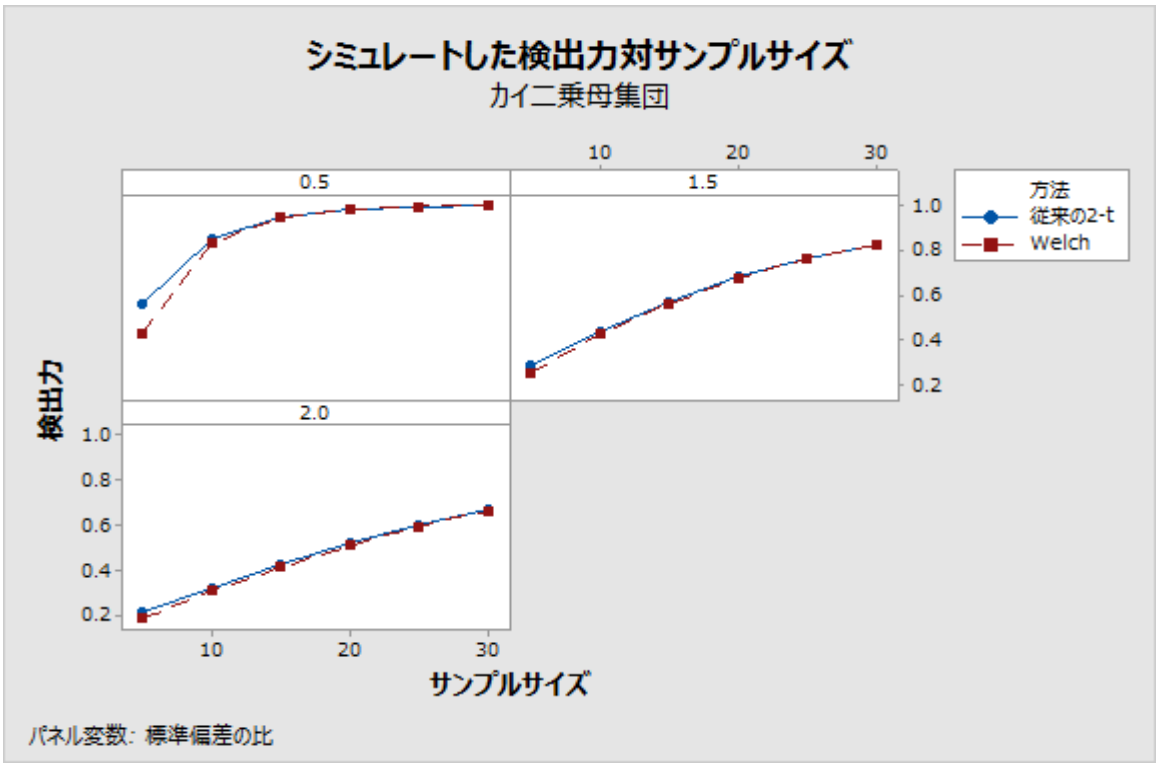

図 6　バランス型不等分散計画での通常の 2 サンプル t 検定と Welch の 2 サンプル t 検定のシミュレートした検出力水準の比較。サンプルは、標準偏差の比が 0.5、1.5、2.0 になるように、分散が等しくないカイ二乗分布から抽出されました。

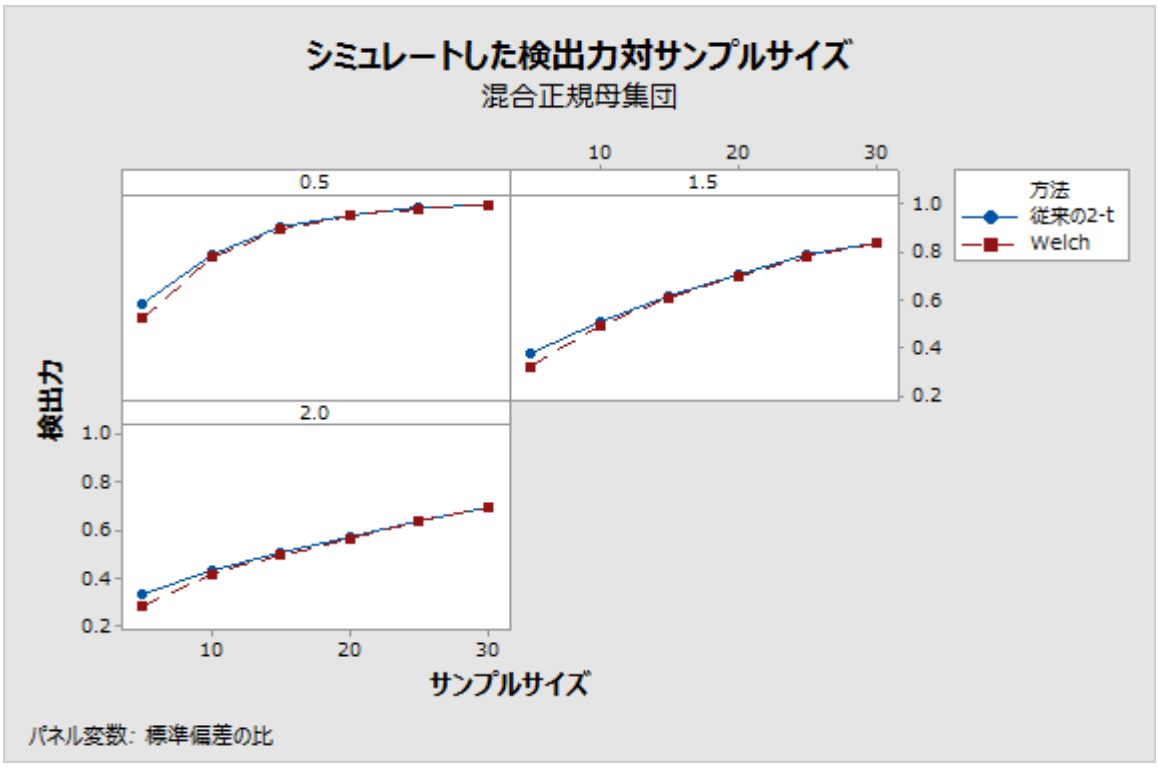

図 7　バランス型不等分散計画での通常の 2 サンプル t 検定と Welch の 2 サンプル t 検定のシミュレートした検出力水準の比較。サンプルは、標準偏差の比が 0.5、1.5、2.0 になるように、分散が等しくない混合正規母集団から抽出されました。

付録 C: 検出力とサンプルサイズおよび正規性に対する感度

アシスタントでは、2 つの母集団の平均を比較する検出力分析は、Welch の t 検定の検出力関数に基づいています。この関数が、正規性の仮定による影響を受ける場合、検出力分析で誤った結論が出る可能性があります。このことから、正規性の仮定に対するこの関数の感度を調べるためにシミュレーションの調査を行いました。サンプルが非正規分布から生成される場合に、感度は、シミュレートした検出力水準と、理論上の検出力関数から計算される検出力水準の間の一貫性として評価されます。定理 B2 によれば、シミュレートした検出力水準と理論上の検出力水準は、サンプルが正規母集団から生成される場合に最も近くなるため、正規分布が基本母集団となります。

シミュレーションの調査 C

この調査は、正規、カイ二乗、混合正規という 3 つの分布を使用して、3 つに分けて行われます。詳細は、「付録 A」を参照してください。調査の各部で、シミュレートした検出力は（与えられたサンプルサイズn_1と与えられた検出可能な差 δ で）、帰無仮説が正しくないときに棄却されるケースの比率として計算されます。すべてのケースで、検出される差は基本母集団の標準の 1 単位で指定されます。これは、この調査の 3 つの分布グループすべてで$\delta = 1.0 \times \sigma_1 = 2.0$になります。比較のために、Welch の t 検定に基づく理論上の検出力も計算されます。

シミュレーション結果と要約

結果は、比較的小さいサンプルサイズの場合、Welch の t 検定の検出力関数は正規性の仮定に対して頑健であることを示しています。一般に、2 つのサンプルの最小サイズが 15 のとき、シミュレートした検出力は対応する目標の理論上の検出力水準に近くなります（表 7～10 および図 8～10 を参照）。

表 7～10 は、正規母集団、歪んだ母集団（カイ二乗）、混合正規母集団から生成されたサンプルペアに基づく$\alpha = 0.05$を使用した両側の Welch の t 検定のシミュレートした検出力水準を示しています。サンプルのペアは同じ分布族からのものですが、親母集団の分散は必ずしも等しくありません。比較のために、理論上の検出力も計算されました。

表 7　n = 5 で$\alpha = 0.05$を使用した両側の Welch の t 検定のシミュレートした検出力水準

			基本母集団: N(0,2)				基本母集団: Chi(2)				基本母集団: 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	観測値	.288	.158	.113	.091	.432	.305	.211	.149	.361	.257	.234	.220

			基本母集団：N(0,2)				基本母集団：Chi(2)				基本母集団：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	$\frac{n_2}{n_1}$		$n_1 = 5$				$n_1 = 5$				$n_1 = 5$			
		目標値	.353	.192	.116	.092	.353	.192	.116	.092	.353	.192	.116	.092
5	1.0	観測値	.370	.252	.169	.121	.427	.334	.248	.189	.522	.380	.319	.284
		目標値	.389	.286	.190	.137	.389	.286	.190	.137	.389	.286	.190	.137
8	1.6	観測値	.387	.326	.242	.179	.427	.364	.286	.225	.573	.453	.374	.319
		目標値	.400	.345	.260	.193	.400	.345	.260	.193	.400	.345	.260	.193
10	2.0	観測値	.390	.351	.272	.208	.421	.373	.296	.235	.590	.483	.394	.336
		目標値	.402	.364	.291	.223	.402	.364	.291	.223	.402	.364	.291	.223

表 8　n = 10 で$\alpha = 0.05$を使用した両側の Welch の t 検定のシミュレートした検出力水準

			基本母集団：N(0,2)				基本母集団：Chi(2)				基本母集団：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	観測値	.651	.346	.197	.131	.768	.493	.320	.221	.689	.484	.404	.358
		目標値	.666	.364	.206	.139	.666	.364	.206	.139	.666	.364	.206	.139
10	1.0	観測値	.742	.556	.369	.254	.831	.612	.430	.308	.776	.619	.496	.419
		目標値	.745	.562	.337	.259	.745	.562	.337	.259	.745	.562	.337	.259
15	1.5	観測値	.765	.641	.483	.358	.865	.679	.511	.377	.792	.679	.547	.456
		目標値	.767	.643	.483	.352	.767	.643	.483	.352	.767	.643	.483	.352

			基本母集団：N(0,2)				基本母集団：Chi(2)				基本母集団：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	$\frac{n_2}{n_1}$		$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
20	2	観測値	.774	.683	.549	.417	.898	.737	.565	.448	.797	.716	.596	.490
		目標値	.777	.686	.551	.422	.777	.686	.551	.422	.777	.686	.551	.422

表 9　n = 15 で$\alpha = 0.05$を使用した両側の Welch の t 検定のシミュレートした検出力水準

			基本母集団：N(0,2)				基本母集団：Chi(2)				基本母集団：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	観測値	.857	.569	.342	.229	.871	.651	.421	.293	.853	.632	.505	.428
		目標値	.861	.568	.338	.221	.861	.568	.338	.221	.861	.568	.338	.221
15	1.0	観測値	.906	.745	.535	.368	.942	.763	.563	.415	.891	.760	.611	.500
		目標値	.910	.753	.541	.379	.910	.753	.541	.379	.910	.753	.541	.379
23	1.53	観測値	.928	.831	.667	.502	.975	.858	.676	.517	.898	.825	.698	.572
		目標値	.925	.830	.670	.509	.925	.830	.670	.509	.925	.830	.670	.509
30	2.0	観測値	.933	.861	.737	.589	.984	.903	.750	.598	.902	.847	.742	.619
		目標値	.931	.863	.736	.589	.931	.863	.736	.589	.931	.863	.736	.589

表 10　n = 20 で$\alpha = 0.05$を使用した両側の Welch の t 検定のシミュレートした検出力水準

			基本母集団: N(0,2)				基本母集団: Chi(2)				基本母集団: 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	観測値	.938	.687	.426	.275	.920	.698	.486	.333	.923	.716	.568	.476
		目標値	.941	.686	.424	.277	.941	.686	.424	.277	.941	.686	.424	.277
20	1.0	観測値	.971	.866	.672	.485	.981	.858	.670	.506	.952	.856	.696	.567
		目標値	.971	.869	.673	.489	.971	.869	.673	.489	.971	.869	.673	.489
30	1.5	観測値	.977	.923	.791	.629	.995	.932	.785	.631	.960	.908	.798	.662
		目標値	.978	.922	.791	.628	.978	.922	.791	.628	.978	.922	.791	.628
40	2.0	観測値	.983	.950	.858	.724	.998	.966	.864	.726	.958	.929	.845	.725
		目標値	.981	.945	.854	.719	.981	.945	.854	.719	.981	.945	.854	.719

2 つのサンプルが正規母集団から生成される場合、サンプルが非常に小さいときでも、シミュレートした検出力は理論上の検出力と一致しています。図 7 に示すように、理論上の検出力とシミュレートした検出力の曲線は、実質的に区別できません。これらの結果は、定理 B2 と一致しています。

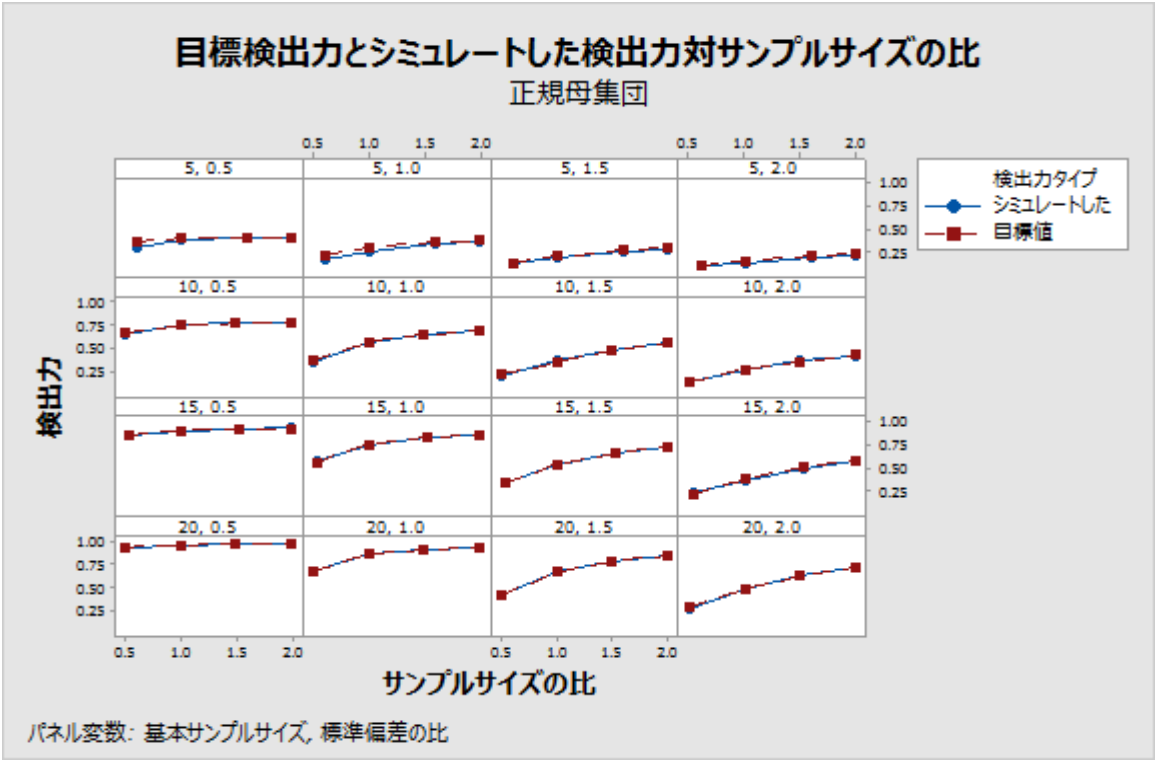

図 8　等分散または不等分散がサンプルサイズ比に対しプロットされた 2 つの正規母集団から生成された一対のサンプルに基づく$\alpha = 0.05$を使用した両側の Welch の t 検定のシミュレートした理論上の検出力水準と目標とする理論上の検出力水準。

サンプルが歪んだカイ二乗分布から生成される場合、非常に小さいサンプルでは、シミュレートした検出力は理論上の検出力より高くなりますが、サンプルサイズが増加するにつれ、検出力は近づきます。図 9 は、2 つのサンプルの最小サイズが 10 以上のとき、目標とする理論上の検出力とシミュレートした検出力の曲線が一貫して近いことを示しています。これは、サンプルが比較的小さい場合でも、Welch の t 検定の検出力関数に対し歪んだデータによる目立った影響はないことを示しています。

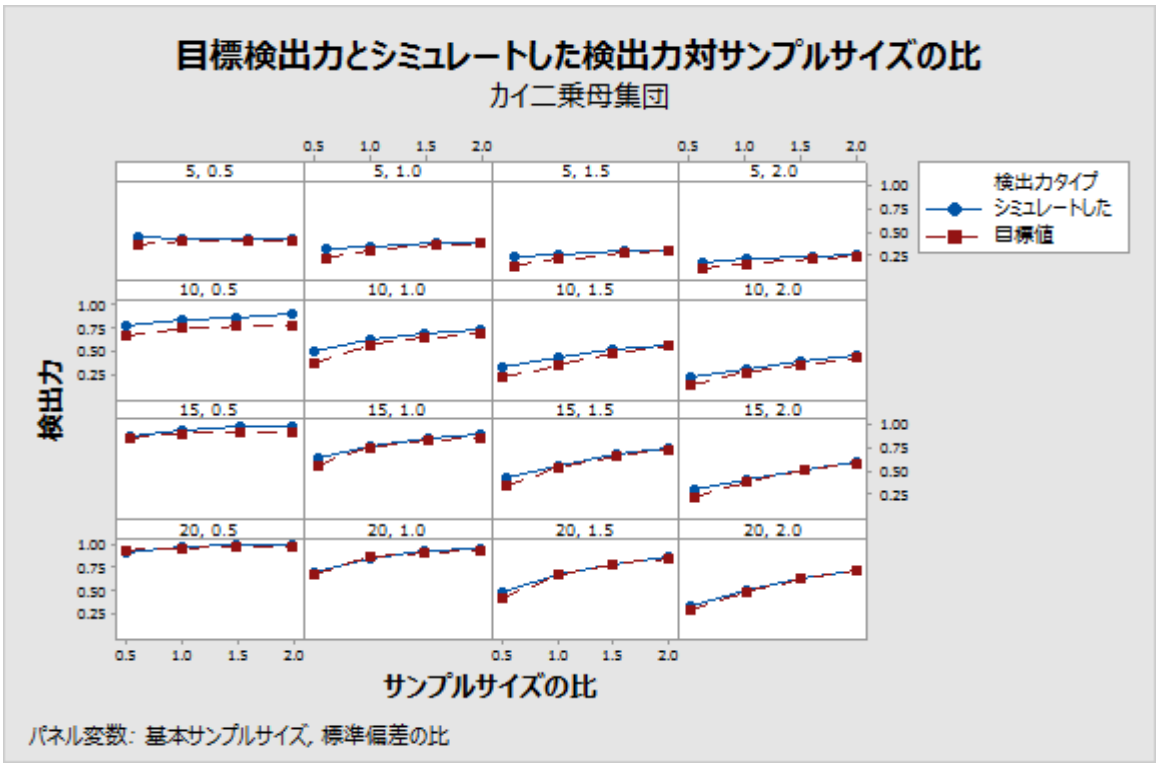

図 9　等分散または不等分散がサンプルサイズ比に対しプロットされた 2 つの正規母集団から生成された一対のサンプルに基づく$\alpha = 0.05$を使用した両側の Welch の t 検定のシミュレートした理論上の検出力水準と目標とする理論上の検出力水準。

さらに、外れ値は、サンプルサイズが非常に小さい場合にのみ、検出力関数に影響を与える傾向があります。一般に、外れ値がある場合、シミュレートした検出力は目標とする理論上の検出力より少し高くなる傾向があります。これは、図 10 に示されています。最小サンプルサイズが 15 に達するまでは、シミュレートした検出力曲線と理論検出力曲線は離れており近接しません。

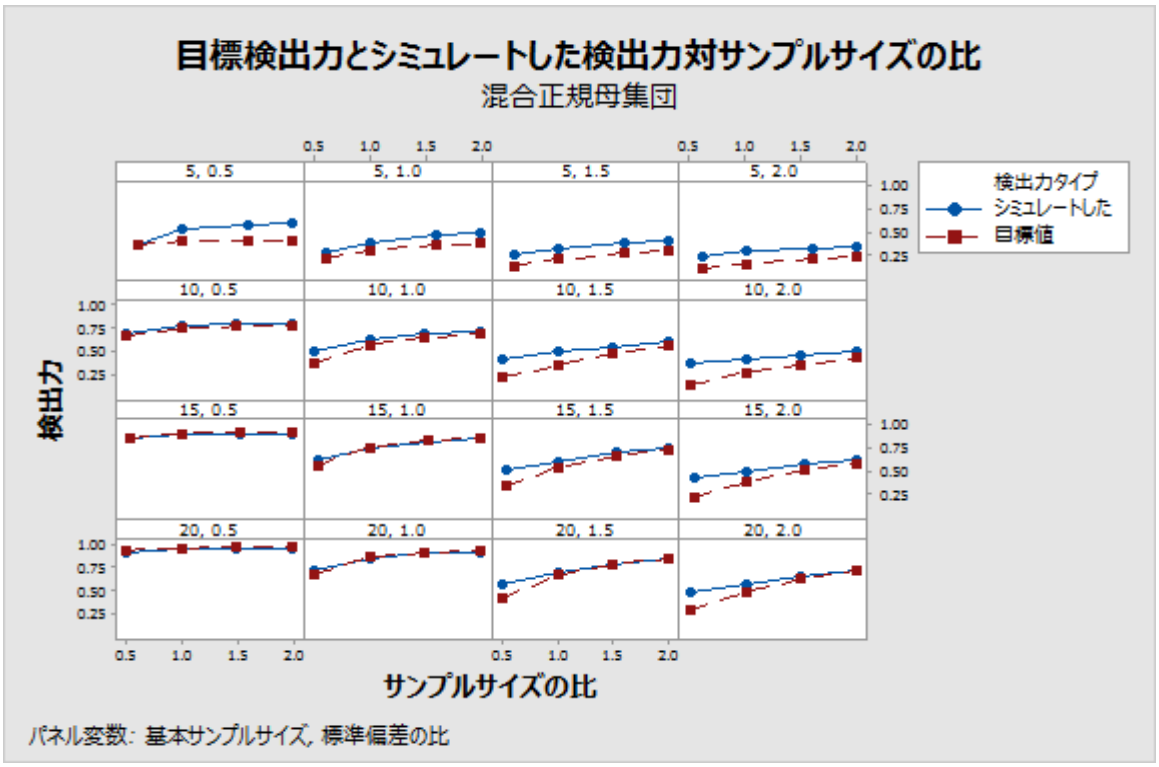

図 10　等分散または不等分散がサンプルサイズ比に対しプロットされた 2 つの正規母集団から生成されたサンプルペアに基づく$\alpha = 0.05$を使用した両側の Welch の t 検定のシミュレートした理論上の検出力水準と目標とする理論上の検出力水準。

付録 D: 定理 B2 の証明

2 サンプルモデルで、帰無仮説のもとに検定統計量

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

を導き出す Welch の手法は、カイ二乗分布に比例する

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

の分布の近似に基づいています。より具体的に言えば、

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

は自由度がd_Wのカイ二乗分布として近似的に分布されます。

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

（1 サンプル設定では、これは$(n-1)s^2/\sigma^2 \sim \chi_{n-1}^2$となる、よく知られた通常の結果です）

対立仮説（つまり$\delta \neq 0$）に対する帰無仮説$H_o: \mu_1 = \mu_2$（つまり$\delta = 0$）の検定を検討します。

帰無仮説では、検出力関数は

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

となり、ここでt_d^αは自由度dの t 分布の 100 α上位百分位数点を意味します。

対立仮説では、

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

は、自由度d_Wと非心パラメータ

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

を持つ近似の非心 t 分布に従います。なぜなら、前述したように

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

は自由度がd_Wのカイ二乗分布として近似的に分布され、

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

は標準正規分布として分布されるからです。

対立仮説では、

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

となります。ここで$G_{d_W,\lambda}(.)$は、上で与えられた自由度d_Wおよび非心パラメータλを使用した非心 t 分布の累積分布関数（C.D.F）です。

付録 E: 定理 B3 の証明

まず最初に、d_Wは次のように書き換えることができます。

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

ここで、$\rho = \sigma_1/\sigma_2$となります。

同様に、Welch の t 検定の検出力関数に関連付けられた非心パラメータも、次のように書き換えることができます。

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

分散が等しいと仮定した場合、通常の 2 サンプル t 検定と Welch 検定に関連付けられた非心パラメータは一致します。つまり

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

ここで、σは 2 つの母集団の共通分散です。したがって、2 つの検定の検出力関数の唯一の差は、それぞれの自由度の差にあります。ただし、分散が等しいと仮定した場合、Welch の t 検定の検出力関数に関連付けられた自由度は、次のようになります。

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

定理 1 によると、通常の 2 サンプル t 検定の検出力関数に関連する自由度は$d_C = n_1 + n_2 - 2$です。いくつか代数的操作を行うと、次のようになります。

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

分散が等しいという仮定のもとでは、通常の 2 サンプル t 検定は UMP（一様最強力）であることがわかっているため、$d - d_W \geq 0$は当然の事実であり、その結果、検出力関数に関連付けられた自由度はより高くなることが予測されます。

そこで、$n_1 \sim n_2$の場合$d \sim d_W$となり、その結果、検出力関数の指標は同じになります。特に、2 つの検定の検出力関数は、$n_1 = n_2$の場合、同一になります。これで、定理 2.3 の最初の部分が証明されます。

$n_1 \neq n_2$の場合、$d_C - d_W > 0$となり、Welch の t 検定は通常の 2 サンプル t 検定より劣ります。

さらに、サンプルが大きい場合、つまり$n_1 \to \infty$および$n_2 \to \infty$の場合、$d_C \to \infty$および$d_W \to \infty$になり、両方の検定に関連付けられた検定統計量の漸近分布は標準正規分布になります。したがって、検定は漸近的に等しく、同じ漸近的検出力関数をもたらします。