



一元配置分散分析 (ANOVA)

概要

一元配置分散分析は、3 つ以上のグループの平均を比較し、互いに有意に異なるかどうかを判断するために使用されます。もう 1 つ重要な機能は、特定のグループ間の差を推定することです。

一元配置分散分析でグループ間の差を検出する最も一般的な方法は F 検定です。これは、すべてのサンプルの母集団が共通だが未知の標準偏差を共有するという仮定に基づきます。実際には、サンプルの標準偏差が異なる場合が多いことは認識しています。したがって、F 検定の代替法として、等しくない標準偏差に対応できる Welch 法について調べる必要がありました。また、標準偏差が等しくないサンプルを考慮する多重比較を計算するための方法を開発する必要がありました。この方法を使用すると、個別区間をグラフ化でき、互いに異なるグループを簡単に特定できます。

本書では、次について、Minitab アシスタントの一元配置分散分析手順で使用される方法をどのように開発したのかについて説明します。

- Welch 検定
- 多重比較区間

さらに、異常なデータの有無、検定のサンプルサイズおよび検出力、データの正規性など、一元配置分散分析の結果の妥当性に影響する可能性がある条件について調べます。これらの条件に基づき、アシスタントでは、データに対して次のチェックが自動的に実行され、レポートカードで結果が報告されます。

- 異常なデータ
- サンプルサイズ
- データの正規性

本書では、これらの条件が一元配置分散分析に実際どのように関連するのかを調べ、アシスタントでこれらの条件をチェックするためのガイドラインをどのように確立したのかについて説明します。

一元配置分散分析法

F 検定と Welch 検定

一元配置分散分析で一般的に使用される F 検定は、すべてのグループが共通だが未知の標準偏差 (σ) を共有するという仮定に基づきます。実際には、この仮定が当てはまることはまれで、その結果、タイプ I 過誤率の制御が難しくなります。タイプ I の誤りとは、帰無仮説を誤って棄却する（サンプルが有意に異ならないのに、異なると結論付ける）確率を指します。サンプルの標準偏差が異なる場合、検定で誤った結論に達する可能性がより高くなります。この問題に対処するために、F 検定の代替法として、Welch 検定が開発されました (Welch 1951)。

目的

アシスタントの一元配置分散分析手順に、F 検定と Welch 検定のどちらを使用するかを判断する必要がありました。そのために、F 検定と Welch 検定の実際の検定結果が、検定の目標有意水準 (α 、すなわちタイプ I 過誤率) とどれくらい厳密に一致するのか、つまり所与のさまざまなサンプルサイズとサンプル標準偏差を使用した場合に、検定で帰無仮説が誤って棄却される回数が目的より多かったか、または少なかったかを評価する必要がありました。

方法

F 検定と Welch 検定を比較するために、さまざまなサンプル数、サンプルサイズ、およびサンプル標準偏差を使用して、複数のシミュレーションを行いました。各条件で、F 検定と Welch 法の両方を使用して分散分析検定を 10,000 回実行しました。サンプルの平均が同じになるように、したがって各検定で帰無仮説が真となるように、ランダムデータを生成しました。次に、0.05 と 0.01 の目標有意水準を使用して検定を実行しました。10,000 回の検定のうち、F 検定と Welch 検定で実際に帰無仮説が棄却された回数を数え、この比率と目標有意水準を比較しました。検定が正しく行われれば、推定されるタイプ I 過誤率は目標有意水準にかなり近くなります。

結果

検定したすべての条件において、Welch 法は F 検定と同等以上の性能を示しました。たとえば、Welch 検定を使用して 5 つのサンプルを比較したとき、タイプ I 過誤率は 0.0460～0.0540 で、目標有意水準の 0.05 にかなり近くなりました。これは、サンプルサイズと標準偏差が全サンプルで異なる場合でも、Welch 法のタイプ I 過誤率は目標値に一致することを示しています。

一方、F 検定のタイプ I 過誤率は 0.0273～0.2277 でした。特に次の条件のとき、F 検定は正しく行われませんでした。

- 最大サンプルの標準偏差もまた最大だった場合、タイプ I 過誤率は 0.05 を下回りました。この条件では検定がより保守的になり、サンプルの標準偏差が等しくない場合は単純にサンプルサイズを増加しても、有効な解決策にはならないことを示しています。
- サンプルサイズが等しいけれども、標準偏差が異なる場合、タイプ I 過誤率は 0.05 を上回りました。標準偏差がより大きいサンプルのサイズが他のサンプルより小さ

い場合でも、過誤率は 0.05 より大きくなりました。特に、より小さいサンプルにより大きい標準偏差があると、この検定で帰無仮説が誤って棄却される危険性が大幅に増加します。

シミュレーションの手法と結果についての詳細は、「付録 A」を参照してください。

Welch 法は標準偏差とサンプルのサイズが等しくない場合に正しく行われたため、アシスタントの一元配置分散分析手順では Welch 法を使用します。

比較区間

分散分析検定が統計的に有意で、少なくとも 1 つのサンプル平均が他と異なることを示す場合、分析の次のステップは、どのサンプルが統計的に異なるのかを判断することです。この比較を行う直観的な方法は、信頼区間をグラフ化し、区間が重なり合っていないサンプルを特定することです。ただし、個別信頼区間は比較するためのものではないため、グラフから引き出す結論が検定結果と一致しない場合があります。標準偏差が等しいサンプルに使用する多重比較方法が発表されていますが、標準偏差が等しくないサンプルを考慮するために、この方法を拡張する必要がありました。

目的

サンプル全体を比較するために使用でき、また検定結果にできるだけ厳密に一致する個別比較区間を計算する方法を開発する必要がありました。また、他のサンプルと統計的に異なるサンプルを判断する視覚的方法を提供する必要もありました。

方法

標準の多重比較の方法 (Hsu 1996) は、多重比較で生じる誤差の増加を制御しながら、平均の各ペア間の差に区間を提供します。サンプルサイズが等しい特別なケースで、標準偏差が等しいと仮定する場合、すべてのペアの差の区間に正確に対応するように各平均の個別区間を表示することができます。Hochberg, Weiss, and Hart (1982) は、サンプルサイズが等しくないケースで、標準偏差が等しいと仮定し、Tukey-Kramer の多重比較の方法に基づいて、ペア間の差の区間にほぼ等しくなる個別区間を開発しました。アシスタントでは、標準偏差が等しいと仮定しない、Games-Howell の多重比較の方法に同じ手法を応用しました。Minitab リリース 16 のアシスタントで使用される手法は、概念的に似ていますが、Games-Howell の手法に直接基づくものではありませんでした。詳細は、「付録 B」を参照してください。

結果

アシスタントの一元配置分散分析要約レポートの平均比較管理チャートに比較区間が表示されます。分散分析検定が統計的に有意である場合、少なくとも他の 1 つの区間と重なり合っていない区間はすべて赤でマークされます。帰無仮説が真のときに棄却される確率は両方の検定で同じなので、そのような結果が出ることはまれですが、検定と比較区間が一致しないことがあります。分散分析検定が有意であるのに、区間のすべてが重なり合っている場合、重なり量の最小のペアが赤でマークされます。分散分析検定が統計的に有意でない場合、重なり合っていない区間があっても、赤でマークされる区間は 1 つもありません。

データチェック

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、特にサンプルが小さい場合には、統計的に有意な結果を検出する確率に影響することがあります。異常なデータは、データ収集での問題を示している可能性や、分析している工程の異常な動作に起因する場合があります。したがって、これらのデータ点は調べる価値があり、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) によって説明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。下位四分位数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。

異常なデータをチェックするときに、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
	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。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

検出力は有意な効果または差が真に存在する場合にそれを識別する尤度を示すため、すべての仮説検定で重要な特性です。検出力は、対立仮説を支持し、帰無仮説を棄却する確率です。多くの場合、検定の検出力を高める最も簡単な方法は、サンプルサイズを大きくすることです。アシスタントでは、検定の検出力が低いと、ユーザーが指定した差を検出するために必要なサンプルの大きさが示されます。差が指定されていない場合は、適度な検出力で検出できる差が表示されます。アシスタントでは Welch 法が使用され、この方法には検出力の厳密式がないため、この情報を提供するためには検出力を計算する方法を開発する必要性がありました。

目的

検出力を計算する手法を開発するため、2 つの問題に取り組む必要性がありました。1 つ目は、アシスタントではユーザーはすべての平均セットを入力する必要がないということです。入力が必要なのは、実用的な意味を持つ平均間の差のみです。どのような差を求める場合でも、その差が生じる可能性のある平均の配置は無数にあります。したがって、可能なすべての平均の配置の検出力を計算することはできないので、検出力を計算するときどの平均を使用するかを判断する妥当な手法を開発する必要性がありました。2 つ目に、アシスタントでは、サンプルサイズまたは標準偏差が等しいことを必要としない Welch 法が使用されるため、検出力を計算する方法を開発する必要性がありました。

方法

無数にある可能な平均の配置に対処するために、Minitab の標準の一元配置分散分析手順（[統計] > [分散分析] > [一元配置]）で使用される手法に基づく方法を開発しました。規定量の差がある平均が 2 つのみで、その他の平均は等しくなるケースに焦点を当てました（平均の加重平均に設定）。全体平均と異なる平均は 2 つのみ（2 つ以下）と仮定するため、推定される検出力は保守的になります。ただし、サンプルのサイズまたは標準偏差が異なる場合があるので、検出力の計算は、どの 2 つの平均が異なると仮定されるのかに依存します。

この問題を解決するために、最良ケースと最悪ケースを表す 2 つの平均ペアを特定します。最悪のケースは、サンプルサイズがサンプル分散に対して比較的小さいときに発生し、検出力は最小化されます。最良ケースは、サンプルサイズがサンプル分散に対して比較的大きいときに発生し、検出力は最大化されます。すべての検出力の計算で、ちょうど 2 つの平均が平均の全体加重平均と異なるという仮定のもとで、検出力が最小化および最大化される 2 つの極端なケースが検討されます。

検出力の計算を開発するために、Kulinskaya et al. (2003) で示された方法を使用しました。シミュレーション、平均の配置に対処するために開発した方法、および Kulinskaya et al. (2003) が示した方法の、それぞれの検出力の計算を比較しました。また、検出力がどのように平均の配置に依存するのかをより明確に示す、別の検出力の近似も調べました。検出力の計算についての詳細は、「付録 C」を参照してください。

結果

これらの方法の比較から、検出力の近似が良好なのは Kulinskaya の方法で、平均の配置を処理する方法は適切だということが示されました。

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

ユーザーが実質的な差を指定すると、この差の最小および最大検出力が計算されます。検出力が90%未満の場合、アシスタントでは指定された差と観測されたサンプル標準偏差に基づいて、サンプルサイズが計算されます。このサンプルサイズで、最小および最大検出力の両方が確実に90%以上となるように、指定された差は、最も大きな変動性を持つ2つの平均の間にあると仮定します。

ユーザーが差を指定しない場合、アシスタントでは、検出力範囲の最大が60%となる最大の差が検出されます。この値は、検出力レポートで、60%の検出力に対応する、赤いバーと黄色のバーの境界にラベル表示されます。また、検出力範囲の最小が90%となる最小の差も検出します。この値は、検出力レポートで、90%の検出力に対応する、黄色のバーと緑のバーの境界にラベル表示されます。

検出力とサンプルサイズを調べる場合、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
	データから平均間に差があると結論付けるのに十分な証拠が得られません。差が指定されていませんでした。
	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。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。
	検出力は十分な可能性があります。検定で平均間の差は検出されませんでした。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。
	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。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。

正規性

多くの統計手法に共通する仮定は、データが正規分布に従っているということです。データが正規分布に従っていない場合でも、正規性の仮定に基づく方法は正しく行われることがあります。これは、すべてのサンプル平均の分布は近似の正規分布に従い、近似はサンプルサイズが大きくなるにつれ、ほぼ正規になるという、中心極限定理である程度説明されます。

目的

目的は、正規分布の適度に良好な近似を得るために必要なサンプルの大きさを判断することでした。さまざまな非正規分布で小規模から中規模のサイズのサンプルを使用して、Welch 検定と比較区間を調べる必要がありました。Welch 法の実際の検定結果と比較区間が、検定の選択された有意水準 (α 、すなわちタイプ I 過誤率) にどれくらい厳密に一致するのか、

つまり所与のさまざまなサンプルサイズ、水準数、非正規分布を使用した場合に、検定で帰無仮説が誤って棄却される回数が期待より多かったか、または少なかったかを判断する必要があります。

方法

タイプ I 過誤率を推定するために、さまざまなサンプル数、サンプルサイズ、およびデータ分布を使用して、複数のシミュレーションを行いました。シミュレーションには、正規分布から大幅に逸脱する歪んだ裾の重い分布が含まれました。サイズおよび標準偏差は、各検定内の全サンプルで一定でした。

各条件で、Welch 法と比較区間を使用して分散分析検定を 10,000 回実行しました。サンプルの平均が同じになるように、したがって各検定で帰無仮説が真となるように、ランダムデータを生成しました。次に、0.05 の目標有意水準を使用して検定を実行しました。10,000 回のうち、検定で実際に帰無仮説が棄却された回数を数え、この比率と目標有意水準を比較しました。比較区間の場合は、10,000 回のうち、区間が差を示した回数を数えました。検定が正しく行われれば、タイプ I 過誤率は目標有意水準に非常に近くなります。

結果

全体として、検定および比較区間は、10 または 15 の小さいサンプルサイズを使用したすべての条件で正しく行われました。水準数が 9 以下の検定の場合、ほとんどすべてのケースで、サンプルサイズ 10 は目標有意水準の 3 パーセントポイント以内、サンプルサイズ 15 は 2 パーセントポイント以内という結果になります。水準数が 10 以上の検定の場合、ほとんどのケースで、サンプルサイズ 15 は目標有意水準の 3 パーセントポイント以内、サンプルサイズ 20 は 2 パーセントポイント以内という結果になります。詳細は、「付録 D」を参照してください。

比較的小さなサンプルで検定が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チェックし、サンプルが水準数 2~9 で 15 未満のとき、水準数 10~12 で 20 未満のときに表示されます。これらの結果に基づき、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
	サンプルサイズが 15 または 20 未満です。正規性に問題はありません。
	一部のサンプルサイズが 15 または 20 未満です。正規性に問題がある可能性があります。

参考文献

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

付録 A: F 検定と Welch 検定

F 検定では、同等標準偏差の仮定に違反すると、タイプ I 過誤率が上昇する可能性があります。Welch 検定はこの問題を回避するために設計されています。

Welch 検定

k 母集団から得たサイズ n_1, \dots, n_k のランダムサンプルが観測されます。 μ_1, \dots, μ_k は母集団平均を表し、 $\sigma_1^2, \dots, \sigma_k^2$ は母集団分散を表すとします。 $\bar{x}_1, \dots, \bar{x}_k$ はサンプル平均を表し、 s_1^2, \dots, s_k^2 はサンプル分散を表すとします。 次の仮説を検定します。

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1: \text{ある } i, j \text{ の } \mu_i \neq \mu_j$$

k 平均の同等性を検定する Welch 検定では、統計量

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

と $F(k-1, f)$ 分布を比較します。ここで、

$$w_j = \frac{n_j}{s_j^2},$$

$$W = \sum_{j=1}^k w_j,$$

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

$$h_j = \frac{(1 - w_j/W)^2}{n_j - 1}, \text{ および}$$

$$f = \frac{k^2 - 1}{3 \sum_{j=1}^k h_j} \text{ となります。}$$

Welch 検定は、 $W^* \geq F_{k-1, f, 1-\alpha}$ (確率 α で超過される F 分布の百分位) の場合、帰無仮説を棄却します。

等しくない標準偏差

このセクションでは、同等標準偏差の仮定の違反に対する F 検定の感度を実証し、Welch 検定と比較します。

次に、 $N(0, \sigma^2)$ の 5 つのサンプルを使用した、一元配置分散分析検定の結果を示します。各行は、F 検定と Welch 検定を使用した 10,000 回のシミュレーションに基づいています。5 番目のサンプルの標準偏差を他のサンプルの 2 倍と 4 倍に増加し、標準偏差について 2 つの条件を検定しました。サンプルサイズについては、サンプルサイズが等しい、5 番目のサンプルが他より大きい、5 番目のサンプルが他より小さい、という 3 つの条件を検定しました。

表1 5つのサンプルと目標有意水準 $\alpha = 0.05$ でシミュレートしたF検定とWelch検定のタイプI過誤率

標準偏差 ($\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5$)	サンプルサイズ (n_1, n_2, n_3, n_4, n_5)	F検定	Welch検定
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

サンプルサイズが等しい場合（行2と行5）、F検定で帰無仮説が誤って棄却される確率は目標の0.05より大きくなり、標準偏差間の非同等性が大きくなると確率が増加します。この問題は、標準偏差が最も大きいサンプルのサイズが減少するとさらに悪化します。一方、標準偏差が最も大きいサンプルのサイズが増加すると、棄却の確率は低下します。ただし、サンプルサイズが増加しすぎると、棄却の確率が低くなりすぎ、帰無仮説での検定が必要以上に保守的になるだけでなく、対立仮説での検定の検出力にも悪影響を及ぼします。すべてのケースで目標有意水準の0.05によく一致する、Welch検定でのこれらの結果を比較します。

次に、 $k = 7$ サンプルを使用するケースのシミュレーションを行いました。表の各行は、10,000回シミュレートしたF検定を要約しています。標準偏差とサンプルサイズはさまざまに変化を持たせました。目標有意水準は、 $\alpha = 0.05$ と $\alpha = 0.01$ です。上記では、目標値からの偏差がかなり大きくなる場合があることがわかります。変動性がより高いときに、より小さいサンプルサイズを使用すると、タイプI過誤率の確率が非常に高くなるのに対し、より大きいサンプルを使用すると、検定が極端に保守的になる可能性があります。次の表2に結果を示します。

表2 7つのサンプルでシミュレートしたF検定のタイプI過誤率

標準偏差 ($\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6, \sigma_7$)	サンプルサイズ ($n_1, n_2, n_3, n_4, n_5, n_6, n_7$)	目標 $\alpha = 0.05$	目標 $\alpha = 0.01$
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

標準偏差 (σ_1 , σ_2 , σ_3 , σ_4 , σ_5 , σ_6 , σ_7)	サンプルサイズ (n_1 , n_2 , n_3 , n_4 , n_5 , n_6 , n_7)	目標 $\alpha =$ 0.05	目標 $\alpha =$ 0.01
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
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

標準偏差 (σ_1 , σ_2 , σ_3 , σ_4 , σ_5 , σ_6 , σ_7)	サンプルサイズ (n_1 , n_2 , n_3 , n_4 , n_5 , n_6 , n_7)	目標 α = 0.05	目標 α = 0.01
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
1.68333、1.68333、1.68333、1.68333、 1.68333、1.68333、3.9	21、21、21、21、22、22、12	0.1340	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

標準偏差 ($\sigma 1$ 、 $\sigma 2$ 、 $\sigma 3$ 、 $\sigma 4$ 、 $\sigma 5$ 、 $\sigma 6$ 、 $\sigma 7$)	サンプルサイズ ($n1$ 、 $n2$ 、 $n3$ 、 $n4$ 、 $n5$ 、 $n6$ 、 $n7$)	目標 $\alpha =$ 0.05	目標 $\alpha =$ 0.01
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
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

付録 B: 比較区間

平均比較管理チャートを使用して、母集団平均間の差の統計的有意性を評価できます。

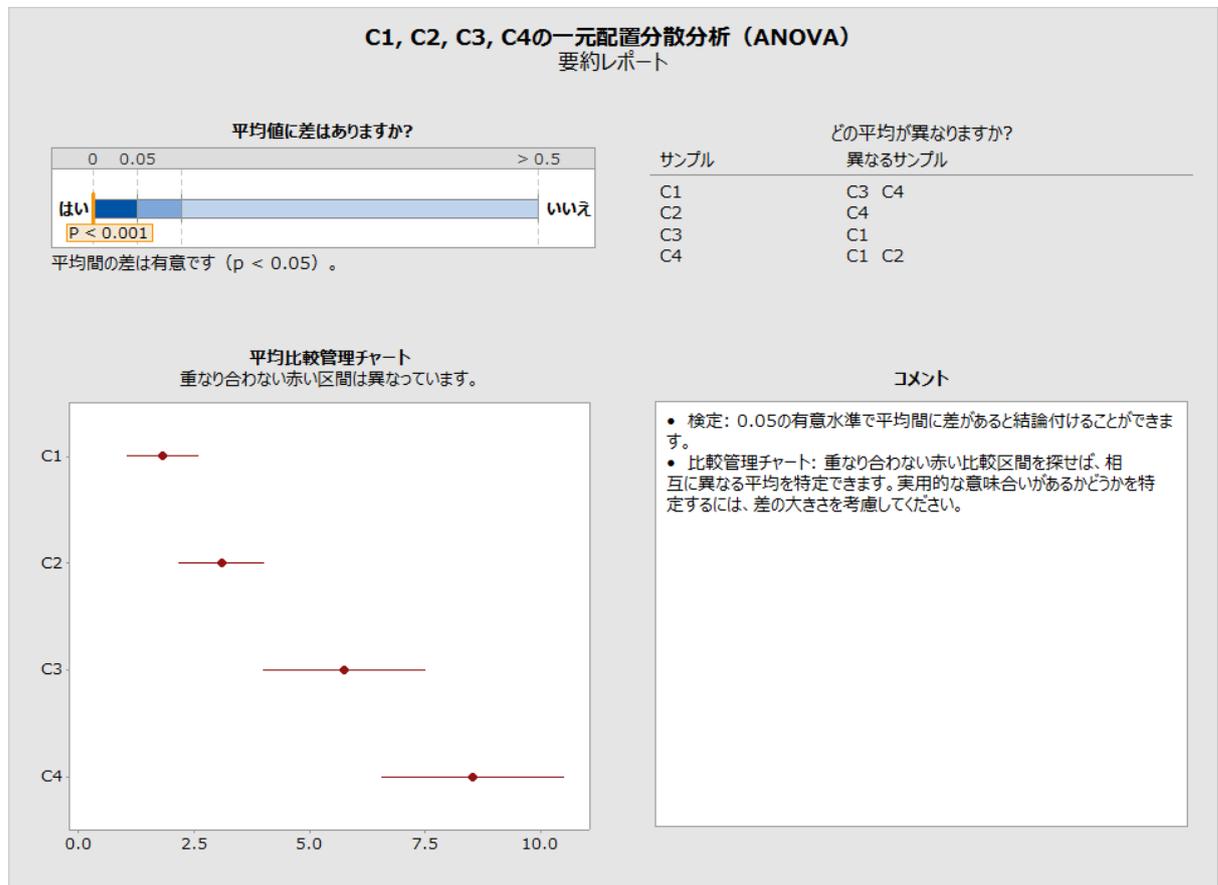
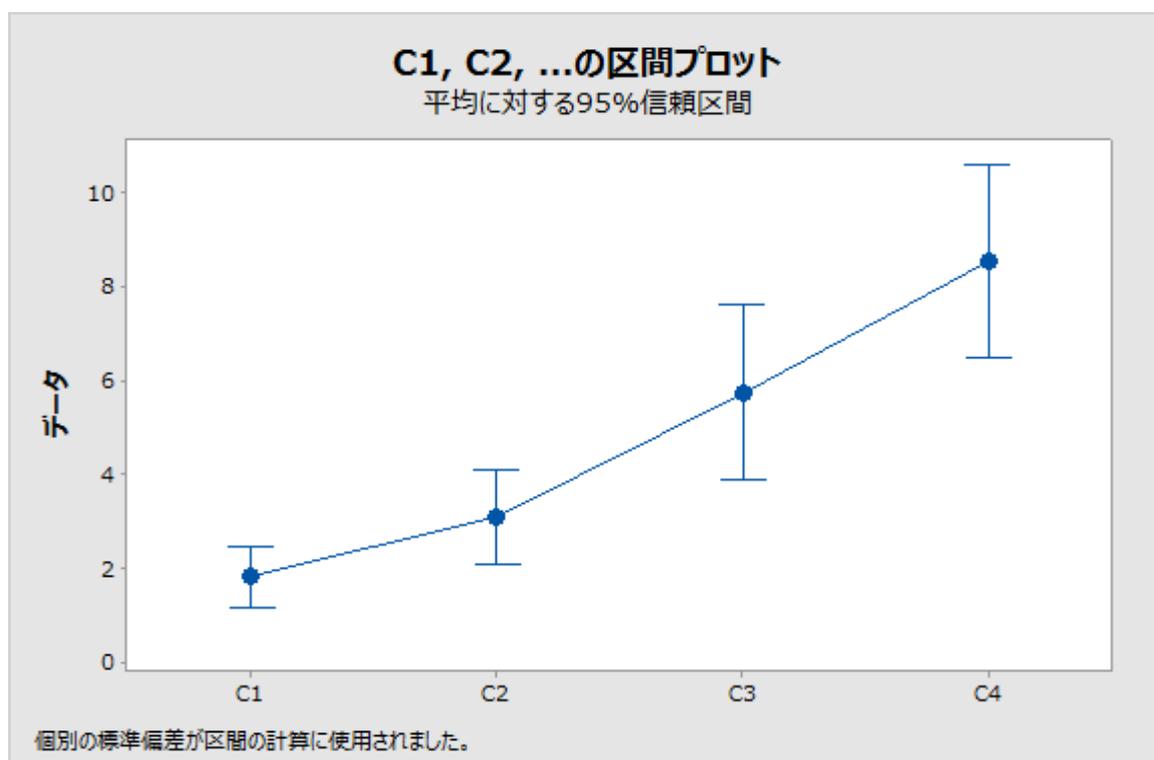


図1 アシスタントの一元配置分散分析要約レポートの平均比較管理チャート

類似の区間セットが、Minitab の標準の一元配置分散分析手順の出力に表示されます（[統計] > [分散分析] > [一元配置]）。



ただし、上記の区間は平均の単なる個別信頼区間です。分散分析検定（FまたはWelchのいずれか）で一部の平均が異なるという結論が出ると、重なり合っていない区間を探し、どの平均が異なるのかについて結論を導こうとする自然な傾向があります。この個別信頼区間の非公式な分析が、妥当な結論につながる場合が多くありますが、誤差の確率は分散分析検定と同様には制御されません。母集団の数によっては、差があると結論付けられる確率が、検定より著しく高くまたは低くなる場合があります。その結果、2つの方法で一致しない結論に容易に達する可能性があります。比較管理チャートは、多重比較を行うときに、Welch 検定の結果により一貫して一致するように設計されています。ただし、完全な一貫性を必ずしも実現できるということではありません。

Minitab の Tukey-Kramer 比較や Games-Howell 比較などの多重比較の方法では（[統計] > [分散分析] > [一元配置]）、個々の平均間の差について統計的に有効な結論を導くことができます。これら2つの方法は、各平均ペアの差の区間が示される、ペアワイズ比較の方法です。推定する差がすべての区間に同時に含まれる確率は少なくとも $1 - \alpha$ です。Tukey-Kramer の方法が等分散性の仮定に依存するのに対し、Games-Howell の方法では等分散性は必要ありません。等平均性の帰無仮説が真の場合、すべての差は0になり、Games-Howell の区間のすべてに0が含まれない確率は、最大でも α です。したがって、有意水準 α で仮説検定を実行するのに、区間を使用できます。アシスタントでは、比較管理チャートの区間を導出するために、Games-Howell の区間を開始点として使用します。

すべての差 $\mu_i - \mu_j$, $1 \leq i < j \leq k$ に区間セット $[L_{ij}, U_{ij}]$ を与え、同じ情報を伝える、個々の平均 μ_i , $1 \leq i \leq k$ の区間セット $[L_i, U_i]$ を求める必要があります。これには、 $\mu_i - \mu_j = d$ となるように $\mu_i \in [L_i, U_i]$ と $\mu_j \in [L_j, U_j]$ が存在する場合にのみ、任意の差 d は区間 $[L_{ij}, U_{ij}]$ にあることが必要です。区間のエンドポイントは次の式に関連する必要があります。

$$U_i - L_j = U_{ij} \text{ と}$$

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

k = 2 の場合、差は 1 つのみですが、個別区間は 2 つあるため、正確な比較区間を得ることができます。実際、この条件を満たす区間の幅にはかなりの柔軟性があります。k = 3 の場合、3 つの差と 3 つの個別区間があるので、条件を満たすことは可能ですが、区間の幅の設定に柔軟性はありません。k = 4 の場合、6 つの差がありますが、個別区間は 4 つしかありません。比較区間では、より少ない区間数で同じ情報を伝えるようにする必要があります。一般に、k ≥ 4 の場合は、個別平均よりも差のほうが多いため、等幅など、差の区間に対して追加条件を科さないかぎり、厳密解はありません。

Tukey-Kramer の区間は、すべてのサンプルサイズが同じ場合にのみ等幅になります。等幅は、等分散性の仮定によって生じる結果でもあります。Games-Howell の区間では、等分散を仮定しないので、等幅にはなりません。アシスタントでは、比較区間を定義するには近似法に頼る必要があります。

$\mu_i - \mu_j$ の Games-Howell の区間は

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

となり、ここで、 $q^*(k, \hat{\nu}_{ij})$ は学生化された範囲分布の近似の百分位数です。これは、比較対象の平均数 k と、

ペア (i, j) に関連付けられた自由度 ν_{ij} に依存します。

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

Hochberg, Weiss, and Hart (1982) は、次の式を使用して、これらのペアワイズ比較にほぼ等しくなる個別区間を求めました。

$$\bar{x}_i \pm |q^*(k, \nu)| s_p X_i$$

値 X_i は、

$$\sum \sum_{i \neq j} (X_i + X_j - a_{ij})^2 \text{ を最小化するために選択されます。}$$

上の式で、

$$a_{ij} = \sqrt{1/n_i + 1/n_j} \text{ となります。}$$

次の形式の Games-Howell の比較から区間を導出して、分散が等しくないケースにこの手法を適応します。

$$\bar{x}_i \pm d_i$$

値 d_i は、

$$\sum \sum_{i \neq j} (d_i + d_j - b_{ij})^2 \text{ を最小化するために選択されます。}$$

上の式で、

$$b_{ij} = |q^*(k, \hat{\nu}_{ij})| \sqrt{s_i^2/n_i + s_j^2/n_j} \text{ となります。}$$

解は

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

次のグラフで、Welch 検定のシミュレーション結果、および2つの方法を使用した比較区間の結果を比較します。この2つの方法とは、現在使用している方法に基づく Games-Howell と、Minitab リリース 16 で使用された自由度の平均に基づく方法です。垂直軸は 10,000 回のシミュレーションのうち、Welch 検定で帰無仮説が誤って棄却された、または比較区間のすべては重なり合っていない回数の比率です。これらの例の目標 α は $\alpha = 0.05$ です。これらのシミュレーションは、標準偏差とサンプルサイズが等しくないさまざまなケースを網羅しています。水平軸に沿った各位置は、異なるケースを表します。

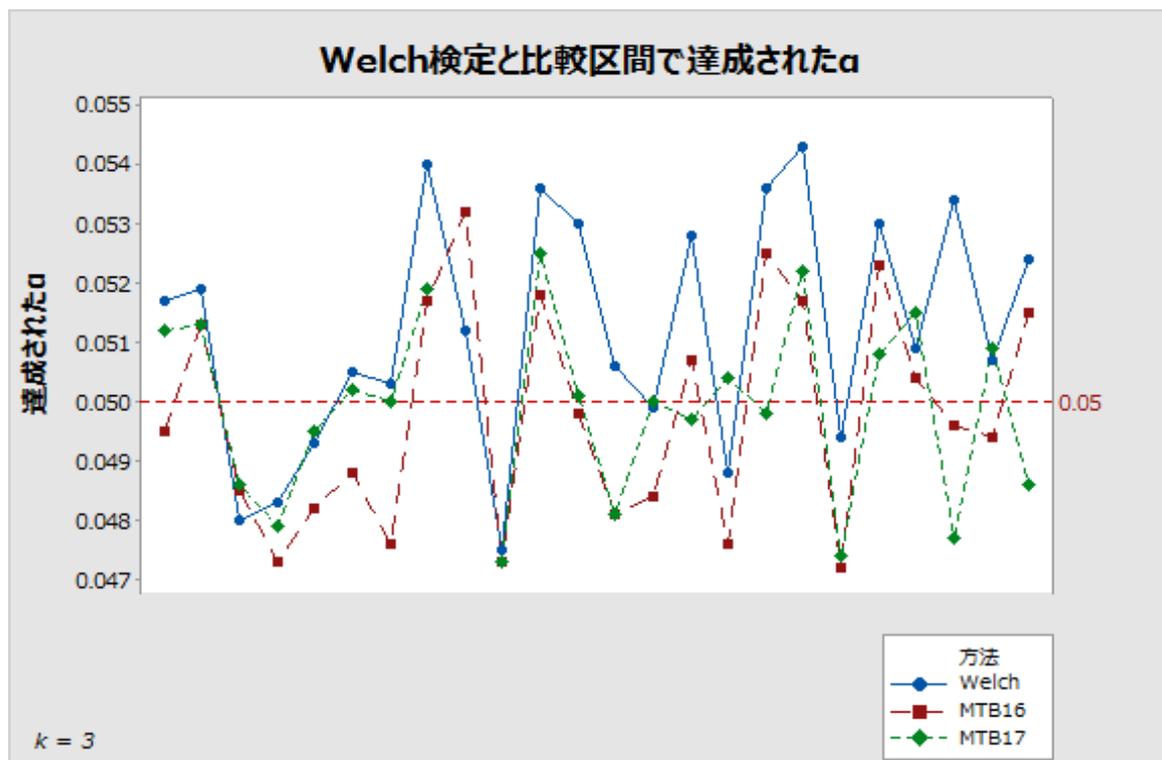


図2 3つのサンプルの比較区間を計算する2つの方法を使用して比較した Welch 検定

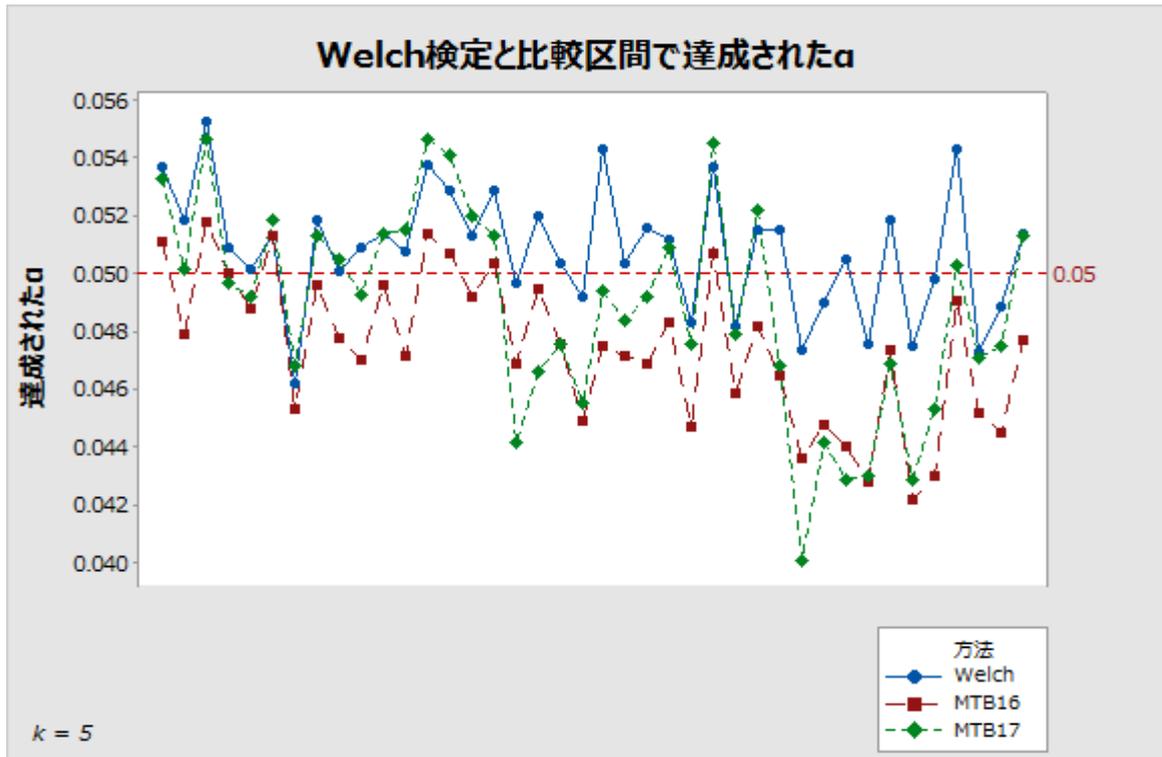


図3 5つのサンプルの比較区間を計算する2つの方法を使用して比較したWelch検定

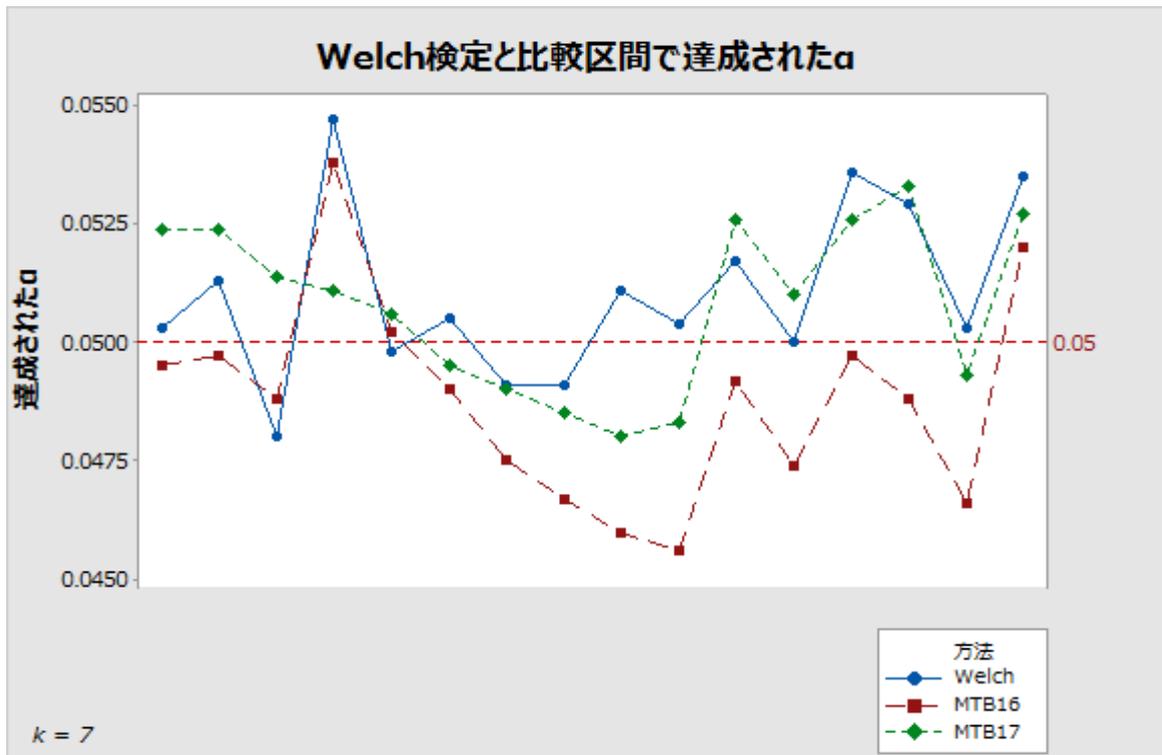


図4 7つのサンプルの比較区間を計算する2つの方法を使用して比較したWelch検定

これらの結果は、目標値の 0.05 周辺の狭い範囲で、シミュレートした α 値を示しています。また、Minitab リリース 17 で実装されている Games-Howell に基づく方法を使用した結果が、

Minitab リリース 16 で使用された方法よりも、Welch 検定の結果に、より厳密に一致することはほぼ間違いありません。

区間の被覆確率が、等しくない標準偏差による影響を受けやすい可能性を示す証拠があります。ただし、この感度は、F 検定の感度ほど極端なものではありません。次のグラフに、 $k = 5$ の場合の、この依存性を示します。

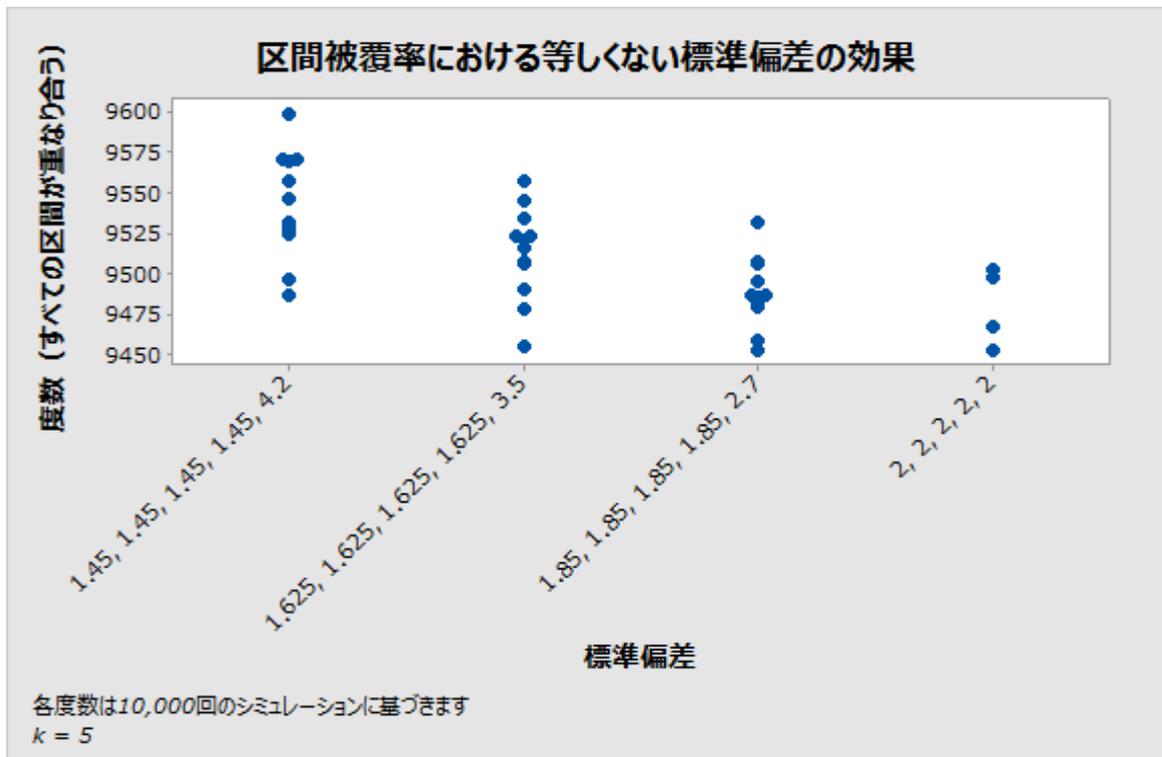


図 5 等しくない標準偏差を使用したシミュレーションの結果

仮説検定と比較区間の併用

まれに、帰無仮説の棄却について、仮説検定と比較が一致しない場合があります。比較区間がすべて重なり合っているのに、検定で帰無仮説が棄却されることがあります。逆に、重なり合っていない区間があるのに、検定で帰無仮説が棄却されないこともあります。真の帰無仮説が棄却される確率は両方の方法で同じなので、この不一致が起こるのはまれです。

これが起こった場合は、最初に検定結果を検討し、比較を使用して有意な検定をさらに調べます。有意水準 α で帰無仮説が棄却される場合、少なくとも他の1つの区間と重なり合っていない比較区間がすべて赤でマークされます。これは、対応するグループ平均が少なくとも他の1つのグループ平均と異なることを視覚的に示すために使用されます。検定が有意で、「最も可能性が高い」差を示す場合、すべての区間が重なり合っていない場合でも、重なり合いの量が最小のペアが赤でマークされます(図6を参照)。特に、重なり合いが非常に小さいペアが他にある場合、これはいくぶんか任意の選択です。ただし、0に近い差に境界があるペアは他にありません。

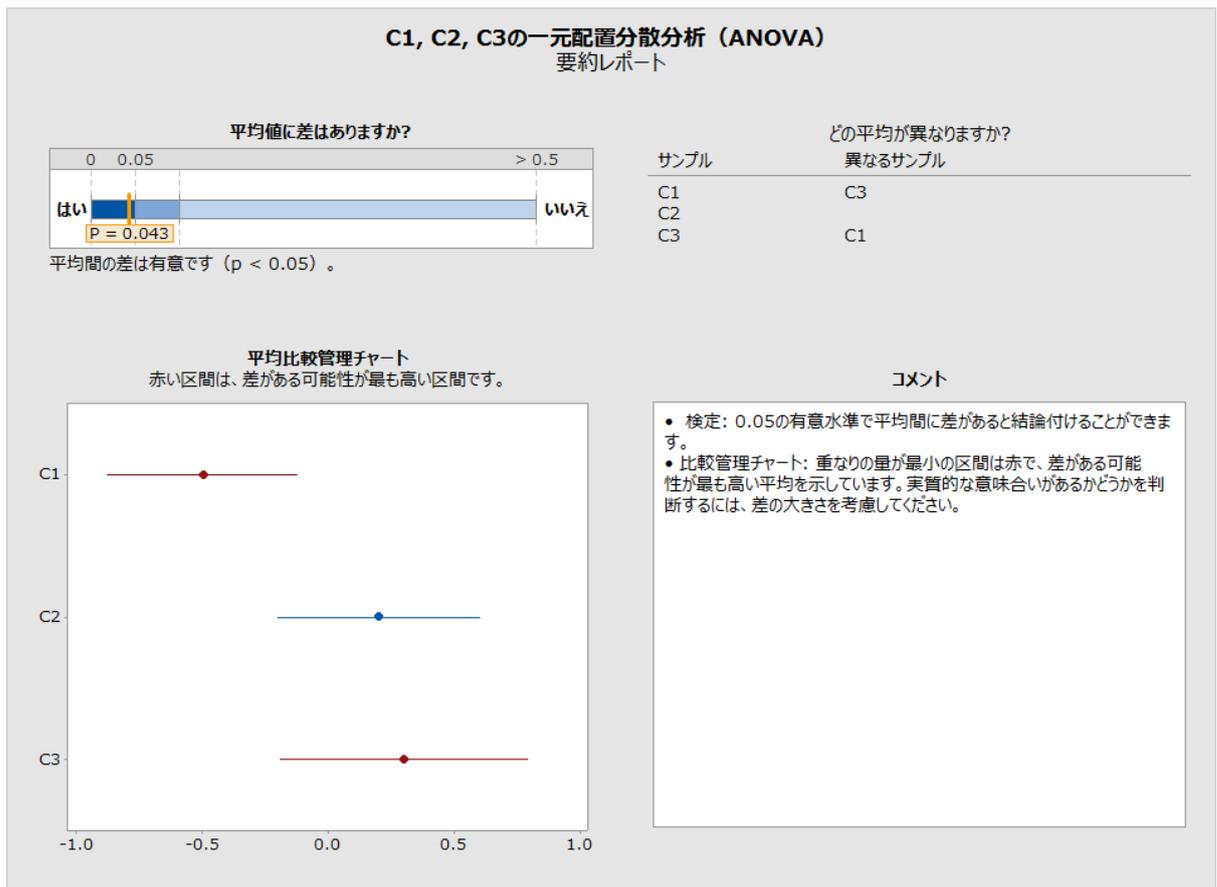


図6 有意な検定。サンプル間で重なり合っている場合でも区間が赤でマークされている

検定で帰無仮説が棄却されない場合、重なり合っていない区間があっても、赤でマークされる区間は1つもありません(次の図7を参照)。区間は平均に差があることを示唆しますが、帰無仮説の棄却の失敗は、帰無仮説が真であると結論付けることと同じではありません。これは単に、観測された差が、原因として可能性を除外するほど大きくないということを示しています。また、この状況では、重なり合っていない区間のGapは通常非常に小さくなるため、非常に小さい差でも区間と一致し、実用的な意味を持つ差があることを必ずしも示しているわけではない点に留意してください。

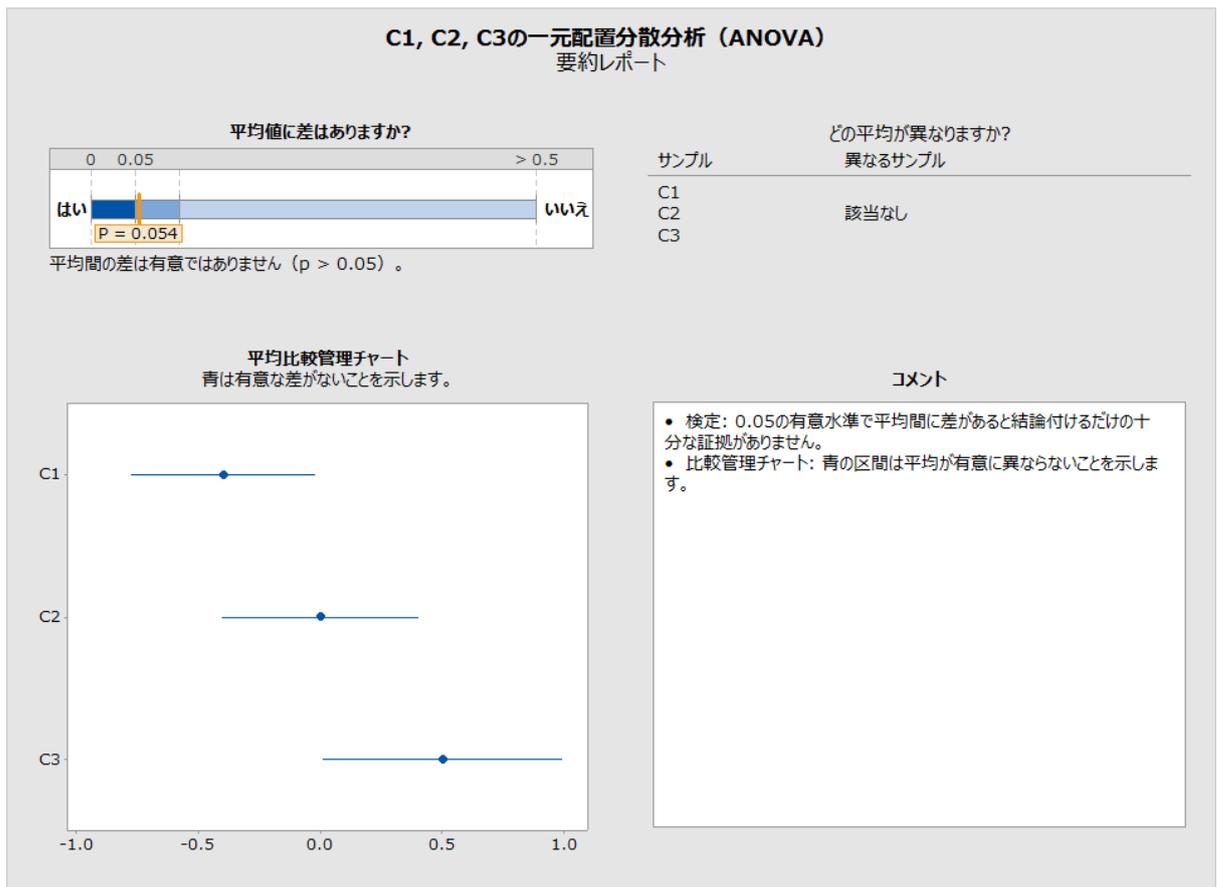


図7 検定失敗。サンプル間の重なり合いがない場合でも区間が赤でマークされていない

付録 C: サンプルサイズ

一元配置分散分析では、検定対象のパラメータはさまざまなグループまたは母集団の母集団平均 $\mu_1, \mu_2, \dots, \mu_k$ です。パラメータがすべて等しい場合、帰無仮説が満たされます。平均間に少しでも差があれば、対立仮説が満たされます。帰無仮説が棄却される確率は、帰無仮説を満たす平均の未満になる必要があります。実際の確率は、分布の標準偏差とサンプルのサイズによって異なります。帰無仮説から偏差を検出する検出力は、より小さい標準偏差またはより大きなサンプルを使用すると増大します。

F 検定の検出力は、非心 F 分布を使用して、標準分散が等しい正規分布の仮定で計算できます。非心パラメータは次のようになります。

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

ここで、 μ は平均の加重平均です。

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

σ は標準偏差で、一定であると仮定します。その他すべてが等しいと、検出力は θ_F で増加します。これは、厳密な意味では、平均が帰無仮説から遠く逸脱するほど、検出力が高くなるということです。

F 検定とは異なり、Welch 検定には検出力の単純な厳密式がありません。そこで、適度に良好な 2 つの近似式を見てみます。最初の式では、F 検定の検出力と類似の方法で非心 F 分布が使用されます。使用する非心パラメータは、次の形式になります。

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

ここで、 μ は加重平均です。

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

ただし重みは、既知の標準偏差 σ_i^2 の結果をシミュレートするのか、サンプル標準偏差 s_i^2 に基づいて検出力を推定するのかによって、標準偏差とサンプルサイズに依存します（つまり $w_i = n_i / \sigma_i^2$ または $w_i = n_i / s_i^2$ ）。近似の検出力は、次のように計算されます。

$$P(F_{k-1, f, \theta_W} \geq F_{k-1, f, 1-\alpha})$$

ここで、分母自由度は

$$f = \frac{k^2 - 1}{3 \sum_{i=1}^k (1 - w_i / \sum_{j=1}^k w_j) / (n_i - 1)}$$
 となります。

次に示すように、これでシミュレーションで観測される検出力に適度に良好な近似が得られます。アシスタントメニューでは検出力の計算に異なる近似を使用しますが、この近似から良い洞察を得ることができ、アシスタントメニューで検出力を計算する平均の配置を選択するための基準となります。

平均の配置

Minitab の検出力とサンプルサイズに使用した手法（[統計] > [分散分析] > [一元配置]）に合わせて、アシスタントでは検出力を評価する平均セットすべての入力には要求されません。代わりに、実用的な意味を持つ平均の差の入力を求められます。最大平均と最小平均に指定

した量の差が生じる可能性のある平均の配置は無数にあります。たとえば、次のすべてで、5つの平均間に10の最大差があります。

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

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

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

この他にも無数の配置があります。

Minitabの検出力およびサンプルサイズで使用される手法（[統計] > [検出力] および [サンプルサイズ] > [一元配置分散分析（ANOVA）]）に従います。つまり、2つを除くすべての平均が平均の（加重）平均となり、残り2つの平均の差が規定量となるケースを抽出します。ただし、分散とサンプルサイズが等しくない可能性があるため、非心パラメータ（したがって検出力）は、どの2つの平均が異なると仮定されるのかによって異なります。

2つを除くすべての平均が全体加重平均 μ に等しく、2つの平均（たとえば $\mu_i > \mu_j$ ）の間、および2つの平均と全体平均の間に差がある、平均の配置 μ_1, \dots, μ_k を検討します。 $\Delta = \mu_i - \mu_j$ が2つの平均間の差を表すとします。 $\Delta_i = \mu_i - \mu$ および $\Delta_j = \mu - \mu_j$ とします。したがって、 $\Delta = \Delta_i + \Delta_j$ となります。また、 μ はすべての k 平均の加重平均を表し、 $(k - 2)$ 個の平均は μ に等しいと仮定されるので、次のようになります。

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

したがって、

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

ゆえに、

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

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

この特定の平均の配置について、Welch検定に関連する非心パラメータを計算できます。

$$\begin{aligned} \theta_W &= w_i(\mu_i - \mu)^2 + w_j(\mu_j - \mu)^2 \\ &= \frac{w_i w_j^2 \Delta^2 + w_j w_i^2 \Delta^2}{(w_i + w_j)^2} = \frac{w_i w_j \Delta^2}{w_i + w_j} \end{aligned}$$

この数量は、固定された w_j の w_i で増加し、逆の場合も同じです。したがって、これは2つの最大の重みを使用したペア(i, j)で最大化され、2つの最小の重みを使用したペアで最小化されます。すべての検出力の計算で、ちょうど2つの平均が平均の全体加重平均と異なるという仮定のもとで、検出力が最小化および最大化される2つの極端なケースが検討されます。

ユーザーが検定の実質的な差を指定すると、この差の最小および最大検出力が評価されます。これらの検出力の範囲は、検出力が60%以下は赤、検出力が90%以上は緑、検出力が60%~90%は黄という色分けしたバーに相対してレポートに示されます。レポートカードの結果は、この色分けされたスケールに相対して、検出力の範囲がどこになるのかによって異なります。

範囲全体が赤の場合、すべてのグループペアの検出力が 60%以下で、レポートカードには検出力不足の問題を示す赤のアイコンが表示されます。範囲全体が緑の場合、すべてのグループの検出力が少なくとも 90%で、レポートカードには十分な検出力の状態を示す緑のアイコンが表示されます。その他すべての状態は、レポートカードに黄色のアイコンで示される、中間の状況として扱われます。

緑の状態に一致しない場合、ユーザーが指定した差と観測されたサンプル標準偏差を所与として、緑の状態がもたらされるサンプルサイズが計算されます。推定される検出力は、重み $w_i = n_i/s_i^2$ を使用するサンプルサイズによって異なります。すべてのサンプルが同じサンプルサイズであると仮定される場合、最も小さい 2 つの重みは、サンプル標準偏差が最も大きい 2 つのグループに対応します。アシスタントでは、指定された差が最も大きな変動性を持つ 2 つのグループの間にある場合に少なくとも 90%の検出力が得られるサンプルサイズが検出されます。したがって、すべてのグループで少なくともこの大きさのサンプルサイズを使用すると、検出力の範囲全体が少なくとも 90%になり、緑の条件が満たされます。

ユーザーが検出力の計算に差を指定しない場合、アシスタントでは、計算された検出力の範囲の最大が 60%となる最大の差が検出されます。この値は、検出力レポートで、60%の検出力に対応する、バーの赤い部分と黄色の部分の境界にラベル表示されます。また、計算された検出力の範囲の最小が 90%となる最小の差も検出されます。この値は、検出力レポートで、90%の検出力に対応する、バーの黄色の部分と緑の部分の境界にラベル表示されます。

検出力の計算

検出力は Kulinskaya et al. (2003) による近似を使用して計算されます。

次を定義します。

$$\begin{aligned}\lambda &= \sum_{i=1}^k w_i (\mu_i - \mu)^2, \\ A &= \sum_{i=1}^k h_i, \\ B &= \sum_{i=1}^k w_i (\mu_i - \mu)^2 (1 - w_i/W)/(n_i - 1), \\ D &= \sum_{i=1}^k w_i^2 (\mu_i - \mu)^4/(n_i - 1), \\ E &= \sum_{i=1}^k w_i^3 (\mu_i - \mu)^6/(n_i - 1)^2\end{aligned}$$

Welch 統計量の分子 $\sum_{i=1}^k w_i (\bar{x}_i - \hat{\mu})^2$ の最初の 3 つのキユムラントは、次のように推定できます。

$$\begin{aligned}\kappa_1 &= k - 1 + \lambda + 2A + 2B, \\ \kappa_2 &= 2(k - 1 + 2\lambda + 7A + 14B + D), \\ \kappa_3 &= 8(k - 1 + 3\lambda + 15A + 45B + 6D + 2E)\end{aligned}$$

$F_{k-1, f, 1-\alpha}$ は、 $F(k-1, f)$ 分布の $(1-\alpha)$ 分位点を表すとし、 $W^* \geq F_{k-1, f, 1-\alpha}$ はサイズ α の Welch 検定で帰無仮説を棄却する基準です。

そこで、

$$\begin{aligned}q &= (k-1) \left[1 + \frac{2(k-2)A}{k^2-1} \right] F_{k-1, f, 1-\alpha}, \\ b &= \kappa_1 - 2\kappa_2^2/\kappa_3,\end{aligned}$$

(注: c の式は Kulinskaya et al. (2003) で括弧なしで示されています) 、

$$\nu = 8\kappa_2^3/\kappa_3^2 \text{ とすると、}$$

Welch 検定の推定される近似の検出力は、次のようになります。

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

ここで、 χ_ν^2 は自由度 ν のカイ二乗確率変数です。

次の結果は、2つの近似法の検出力と、10,000回のシミュレーションに基づく、例の範囲のシミュレートされた検出力を比較したものです。

表3 シミュレートした検出力と比較した2つの近似法の検出力の計算

例	α	シミュレートした検出力	非心 F	Kulinskaya et al.
μ : 0, 0, 0, -0.1724, 0.8276	0.10	0.1372	0.135702	0.135795
σ : 2, 2, 2, 2, 4	0.05	0.0739	0.072563	0.069512
n : 12, 12, 12, 12, 10	0.01	0.0195	0.016587	0.012538
μ : 0, 0, 0, -0.3448, 1.6552	0.10	0.2498	0.251064	0.257455
σ : 2, 2, 2, 2, 4	0.05	0.1574	0.153128	0.156215
n : 12, 12, 12, 12, 10	0.01	0.0541	0.045211	0.042195
μ : 0, 0, 0, -0.5172, 2.4828	0.10	0.4534	0.445570	0.453506
σ : 2, 2, 2, 2, 4	0.05	0.3211	0.311994	0.321575
n : 12, 12, 12, 12, 10	0.01	0.1273	0.121225	0.125065
μ : 0, 0, 0, -0.6896, 3.3104	0.10	0.6620	0.671317	0.670296
σ : 2, 2, 2, 2, 4	0.05	0.5219	0.533819	0.538617
n : 12, 12, 12, 12, 10	0.01	0.2842	0.271316	0.282759
μ : 0, 0, 0, -0.8620, 4.1380	0.10	0.8417	0.852589	0.846697
σ : 2, 2, 2, 2, 4	0.05	0.7382	0.752173	0.746121
n : 12, 12, 12, 12, 10	0.01	0.4883	0.487601	0.493230
μ : 0, 0, 0, -1.0344, 4.9656	0.10	0.9429	0.952077	0.954929
σ : 2, 2, 2, 2, 4	0.05	0.8866	0.901485	0.897937
n : 12, 12, 12, 12, 10	0.01	0.6910	0.711055	0.703379
μ : 0, 0, 0, 0, 0, -0.148148, 1.85185	0.10	0.2011	0.189392	0.200114
σ : 2, 2, 2, 2, 2, 2, 5	0.05	0.1201	0.108986	0.117420
n : 20, 20, 20, 20, 20, 20, 10	0.01	0.0385	0.028986	0.031456

例	α	シミュレートした検出力	非心 F	Kulinskaya et al.
μ : 0、0、0、0、0、-0.296296、3.70370	0.10	0.4942	0.485917	0.500143
σ : 2、2、2、2、2、2、5	0.05	0.3677	0.351593	0.375296
n: 20、20、20、20、20、20、10	0.01	0.1770	0.149041	0.177189
μ : 0、0、0、0、0、-0.444444、5.55556	0.10	0.8125	0.829702	0.819542
σ : 2、2、2、2、2、2、5	0.05	0.7131	0.727384	0.720807
n: 20、20、20、20、20、20、10	0.01	0.4876	0.474291	0.494690
μ : 0、0、0、0、0、-0.592593、7.40741	0.10	0.9645	0.977211	0.984213
σ : 2、2、2、2、2、2、5	0.05	0.9286	0.949997	0.949239
n: 20、20、20、20、20、20、10	0.01	0.7938	0.831174	0.814067
μ : 0、0、0、0、0、-0.740741、9.25926	0.10	0.9961	0.998947	1.00000
σ : 2、2、2、2、2、2、5	0.05	0.9895	0.996653	1.00000
n: 20、20、20、20、20、20、10	0.01	0.9528	0.977536	0.98705
μ : 0、0、0、0、0、-0.888889、11.1111	0.10	0.9999	0.999985	1.00000
σ : 2、2、2、2、2、2、5	0.05	0.9995	0.999926	1.00000
n: 20、20、20、20、20、20、10	0.01	0.9943	0.998910	1.00000
μ : 0、0、0、0、0、-0.518519、6.48148	0.10	0.9059	0.929392	0.924696
σ : 2、2、2、2、2、2、5	0.05	0.8403	0.868721	0.856720
n: 20、20、20、20、20、20、10	0.01	0.6511	0.671210	0.666520
μ : 0、0、0、0、0、-.5、.5	0.10	0.1870	0.186658	0.183290
σ : 2、2、2、2、2、2、2	0.05	0.1098	0.106600	0.100189
n: 12、12、12、12、12、12、12	0.01	0.0315	0.027773	0.021332
μ : 0、0、0、0、0、-1、1	0.10	0.4734	0.474736	0.472469
σ : 2、2、2、2、2、2、2	0.05	0.3394	0.338655	0.334430
n: 12、12、12、12、12、12、12	0.01	0.1378	0.137788	0.128693
μ : 0、0、0、0、0、-1.5、1.5	0.10	0.8228	0.817355	0.810181
σ : 2、2、2、2、2、2、2	0.05	0.7112	0.707319	0.698461
n: 12、12、12、12、12、12、12	0.01	0.4391	0.441154	0.431868
μ : 0、0、0、0、0、-2、2	0.10	0.9691	0.973246	0.973319
σ : 2、2、2、2、2、2、2	0.05	0.9312	0.940585	0.936546
n: 12、12、12、12、12、12、12	0.01	0.7817	0.799339	0.785099

例	α	シミュレートした検出力	非心 F	Kulinskaya et al.
μ : 0、0、0、0、0、-2.5、2.5 σ : 2、2、2、2、2、2、2 n: 12、12、12、12、12、12、12	0.10 0.05 0.01	0.9984 0.9936 0.9587	0.998579 0.995330 0.967674	0.999763 0.997481 0.966249
μ : 0、0、0、0、0、-3、3 σ : 2、2、2、2、2、2、2 n: 12、12、12、12、12、12、12	0.10 0.05 0.01	1.0000 0.9997 0.9959	0.999975 0.999870 0.997927	1.00000 1.00000 0.99961
μ : 0、0、0、0、0、-3.5、3.5 σ : 2、2、2、2、2、2、2 n: 12、12、12、12、12、12、12	0.10 0.05 0.01	1.00000 1.00000 0.99998	1.00000 1.00000 0.99995	1.00000 1.00000 1.00000
μ : 0、0、0、0、0、-1.75、1.75 σ : 2、2、2、2、2、2、2 n: 12、12、12、12、12、12、12	0.10 0.05 0.01	0.9140 0.8418 0.6190	0.921225 0.852755 0.633815	0.916652 0.843856 0.620704
μ : 0、-0.5、0.5 σ : 2、2、2 n: 12、12、12	0.10 0.05 0.01	0.2548 0.1549 0.0470	0.259249 0.160861 0.049045	0.257149 0.156251 0.042292
μ : 0、-1、1 σ : 2、2、2 n: 12、12、12	0.10 0.05 0.01	0.6540 0.5205 0.2612	0.659073 0.522885 0.263550	0.654105 0.515816 0.252469
μ : 0、-1.5、1.5 σ : 2、2、2 n: 12、12、12	0.10 0.05 0.01	0.9364 0.8747 0.6614	0.935939 0.875620 0.664478	0.937768 0.872608 0.652563
μ : 0、-1.75、1.75 σ : 2、2、2 n: 12、12、12	0.10 0.05 0.01	0.9810 0.9522 0.8251	0.981434 0.956100 0.830726	0.986815 0.959796 0.823624
μ : 0、-2、2 σ : 2、2、2 n: 12、12、12	0.10 0.05 0.01	0.9953 0.9878 0.9308	0.995969 0.988175 0.931922	0.999332 0.993705 0.933446
μ : 0、-2.5、2.5 σ : 2、2、2 n: 12、12、12	0.10 0.05 0.01	0.9999 0.9997 0.9949	0.999923 0.999634 0.994725	1.00000 1.00000 0.99909

例	α	シミュレートした検出力	非心 F	Kulinskaya et al.
μ : 0, -3, 3	0.10	1.0000	1.00000	1.00000
σ : 2, 2, 2	0.05	1.0000	1.00000	1.00000
n: 12, 12, 12	0.01	0.9999	0.99985	1.00000
μ : 0, -3.5, 3.5	0.10	1.0000	1.00000	1.00000
σ : 2, 2, 2	0.05	1.0000	1.00000	1.00000
n: 12, 12, 12	0.01	0.9999	1.00000	1.00000
μ : 0, -0.142857, 0.857143	0.10	0.1452	0.143156	0.146824
σ : 2, 2, 4	0.05	0.0790	0.077699	0.077538
n: 14, 12, 8	0.01	0.0223	0.018200	0.014338
μ : 0, -0.285714, 1.71429	0.10	0.2765	0.274240	0.286222
σ : 2, 2, 4	0.05	0.1787	0.170628	0.179469
n: 14, 12, 8	0.01	0.0624	0.051588	0.050335
μ : 0, -0.428571, 2.57143	0.10	0.4861	0.476925	0.490018
σ : 2, 2, 4	0.05	0.3487	0.338626	0.355743
n: 14, 12, 8	0.01	0.1467	0.132405	0.141352
μ : 0, -0.50000, 3	0.10	0.5846	0.588533	0.596795
σ : 2, 2, 4	0.05	0.4425	0.444491	0.460707
n: 14, 12, 8	0.01	0.2107	0.197290	0.212798
μ : 0, -0.571429, 3.42857	0.10	0.6933	0.694684	0.696773
σ : 2, 2, 4	0.05	0.5631	0.555731	0.567129
n: 14, 12, 8	0.01	0.3052	0.279131	0.299302
μ : 0, -0.714286, 4.28571	0.10	0.8480	0.861469	0.859329
σ : 2, 2, 4	0.05	0.7402	0.759703	0.759762
n: 14, 12, 8	0.01	0.4871	0.480052	0.497421
μ : 0, -0.857143, 5.14286	0.10	0.9434	0.952562	0.961913
σ : 2, 2, 4	0.05	0.8869	0.898817	0.902716
n: 14, 12, 8	0.01	0.6649	0.687058	0.692591
μ : 0, -1, 6	0.10	0.9849	0.987981	0.999989
σ : 2, 2, 4	0.05	0.9609	0.967589	0.985049
n: 14, 12, 8	0.01	0.8294	0.847436	0.853787

例	α	シミュレートした検出力	非心 F	Kulinskaya et al.
μ : 0, -1.14286, 6.85714 σ : 2, 2, 4 n: 14, 12, 8	0.10	0.9976	0.997776	1.00000
	0.05	0.9890	0.992220	1.00000
	0.01	0.9222	0.940972	0.96383
μ : 1, 2, 3 σ : 0.3, 2.4, 3.6 n: 13, 19, 25	0.10	0.8838	0.882194	0.884649
	0.05	0.7995	0.797869	0.802137
	0.01	0.5632	0.556486	0.563208
μ : 1, 2, 3 σ : 2.77489, 2.77489, 2.77489 n: 13, 19, 25	0.10	0.5649	0.566831	0.565141
	0.05	0.4305	0.431302	0.428126
	0.01	0.1994	0.201329	0.195734

上の結果を次のグラフに要約します。このグラフは、各近似とシミュレーションによって推定された検出力の値の不一致を示しています。

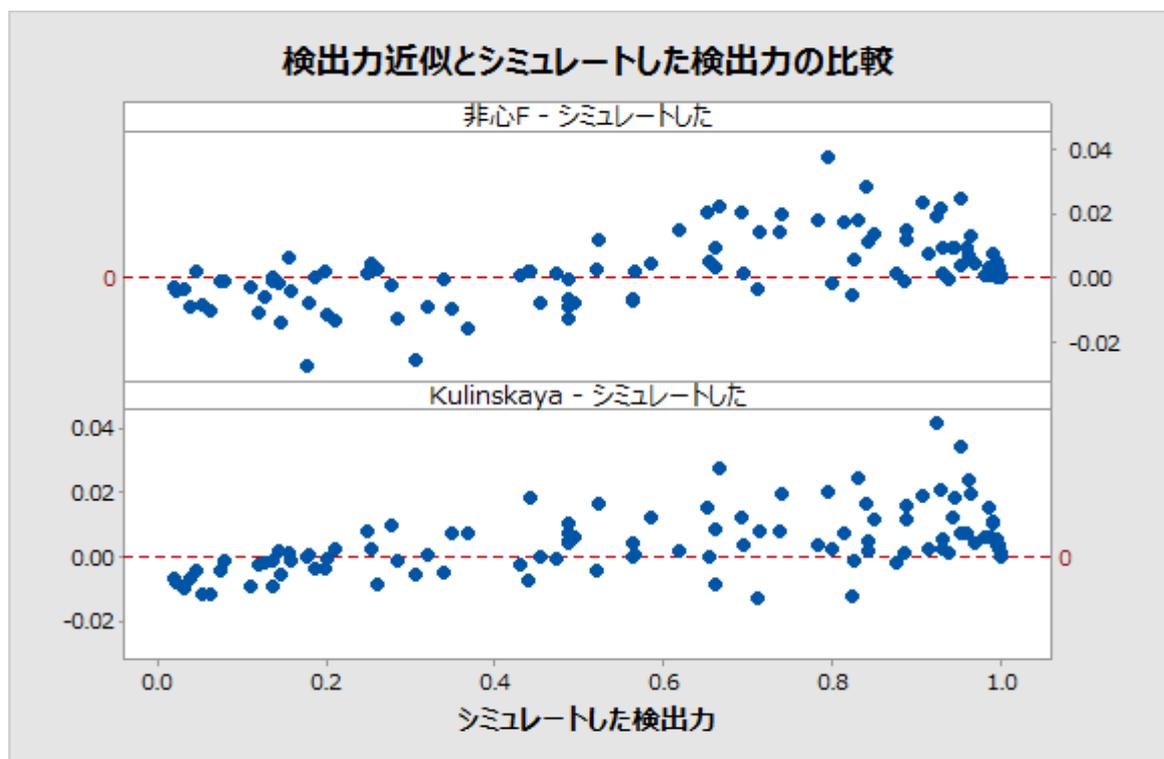


図8 2つの検出力の近似とシミュレーションによって推定された検出力の比較

付録 D: 正規性

このセクションでは、さまざまな非正規分布で小規模から中規模のサンプルを使用して、Welch 検定および比較区間の性能を調べるシミュレーションについて説明します。

次の表に、等平均性の帰無仮説のもとで、さまざまなタイプの分布で行ったシミュレーションの結果を要約します。これらの例では、すべての標準偏差が等しく、すべてのサンプルのサイズも等しくなっています。サンプル数は、 $k = 3, 5$ 、または 7 です。

各セルは、シミュレーション 10,000 回に基づくタイプ I の誤りの推定値を示しています。目標有意水準（目標 α ）は 0.05 です。

表 4 さまざまな分布で等しい平均を使用した Welch 検定のシミュレーション結果

分布	サンプルサイズ $n = 10$			サンプルサイズ $n = 15$		
	$k = 3$	$k = 5$	$k = 7$	$k = 3$	$k = 5$	$k = 7$
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
ラプラス分布 Laplace(0, 1)	0.0433	0.0354	0.0345	0.0445	0.0397	0.0407
一様分布 Uniform(-1, 1)	0.0544	0.0640	0.0718	0.0517	0.0573	0.0585
ベータ分布 Beta(3, 3)	0.0504	0.0577	0.0622	0.0501	0.0538	0.0564
指数分布	0.0508	0.0621	0.0748	0.0483	0.0633	0.0779
カイ二乗分布 Chi-square(3)	0.0473	0.0579	0.0753	0.0499	0.0588	0.0703
カイ二乗分布 Chi-square(5)	0.0458	0.0594	0.0643	0.0504	0.0606	0.0679
カイ二乗分布 Chi-square(10)	0.0463	0.0510	0.0585	0.0463	0.0552	0.0567
ベータ分布 Beta(8, 1)	0.0500	0.0622	0.0775	0.0549	0.0653	0.0760

タイプ I 過誤率は、サンプルのサイズが 10 の場合でも、すべて目標 α の 3 パーセントポイント以内になりました。グループ数が多く、正規から離れた分布では、偏差が大きくなる傾向があります。サンプルサイズが 10 のとき、合格確率が 2 パーセントポイント以上外れるのは、 $k = 7$ でした。これは、正規よりかなり裾が短い一様分布、大きく歪んだ指数分布、カイ二乗分布 Chi-square(3)、ベータ分布 Beta(8, 1) で起こりました。サンプルサイズを 15 まで増加すると、一様分布の結果は著しく向上しますが、 2 つの大きく歪んだ分布では向上しません。

比較区間に対して同様のシミュレーションを行いました。この場合のシミュレートした α は、シミュレーション数 $10,000$ のうち、一部の区間が重なり合っていない数を指します。目標 $\alpha = 0.05$ です。

表5 さまざまな分布で等しい平均を使用した比較区間のシミュレーション結果

分布	サンプルサイズ n = 10			サンプルサイズ n = 15		
	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
ラプラス分布 Laplace(0, 1)	0.0438	0.0305	0.0246	0.0456	0.0366	0.0348
一様分布 Uniform(-1, 1)	0.0559	0.0605	0.0699	0.0534	0.0607	0.0590
ベータ分布 Beta(3, 3)	0.0515	0.0569	0.0615	0.0510	0.0553	0.0568
指数分布	0.0353	0.0254	0.0207	0.0346	0.0310	0.0275
カイ二乗分布 Chi-square(3)	0.0375	0.0305	0.0296	0.0384	0.0359	0.0339
カイ二乗分布 Chi-square(5)	0.0405	0.0390	0.0353	0.0417	0.0433	0.0416
カイ二乗分布 Chi-square(10)	0.0425	0.0428	0.0447	0.0435	0.0476	0.0464
ベータ分布 Beta(8, 1)	0.0381	0.0352	0.0287	0.0459	0.0428	0.0403

Welch 検定と同様に、タイプ I 過誤率は、サンプルのサイズが 10 の場合でも、すべて目標 α の 3 パーセントポイント以内になりました。サンプル数が多く、正規から離れた分布では、偏差が大きくなる傾向があります。サンプルサイズが 10 のとき、 $k = 7$ では過誤率が 2 パーセントポイント以上外れる場合があります ($k = 5$ の場合は 1 件)。このようなケースは、自由度 3 の極端に裾の重い t 分布、ラプラス分布、大きく歪んだ指数分布およびカイ二乗分布 Chi-square(3) で起こります。サンプルサイズを 15 まで増加すると、結果は著しく向上しますが、 $t(3)$ と指数分布のシミュレートした α 値のみが目標から 2 パーセントポイント以上外れます。ただし、Welch 検定の結果とは異なり、比較区間のより大きな偏差は保守的です。

アシスタントの一元配置分散分析では、最大で $k = 12$ のサンプルを使用できるため、次にサンプル数 8 以上の結果を検討します。次の表に、 $k = 9$ グループの非正規データで Welch 検定を使用したタイプ I 過誤率を示します。ここでも、目標 $\alpha = 0.05$ です。

表6 9 個のサンプルを使用したさまざまな分布の Welch 検定のシミュレーション結果

分布	k = 9
t(3)	0.0362
t(5)	0.0426
ラプラス分布 Laplace(0, 1)	0.0402
一様分布 Uniform(-1, 1)	0.0625

ベータ分布 Beta(3, 3)	0.0584
指数分布	0.0885
カイ二乗分布 Chi-square(3)	0.0774
カイ二乗分布 Chi-square(5)	0.0686
カイ二乗分布 Chi-square(10)	0.0581
ベータ分布 Beta(8, 1)	0.0863

予想どおり、大きく歪んだ分布は、目標 α から最も大きな偏差を示しています。その場合でも、指数分布の偏差は近いですが、目標からの逸脱が4パーセントポイントを超える過誤率はありません。すべての結果が目標 α に少なくとも適度に近いため、レポートカードでは、サンプルのサイズ15は十分であるとして扱い、非正規データの問題は示されません。

サイズ $n = 15$ のサンプルは、 $k = 12$ のサンプルになると、それほど正しく行われません。次に、極端に非正規な分布を使用して、さまざまなサンプルサイズで行った Welch 検定のシミュレートした結果を検討します。これは、サンプルサイズの適度な基準を開発するのに役立ちます。

表7 12個のサンプルを使用したさまざまな分布の Welch 検定のシミュレーション結果

n	t(3)	一様分布	カイ二乗分布 Chi-square(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

目標 α からの偏差が2パーセントポイントを若干上回っても許容できるならば、これらの分布で $n = 15$ は許容範囲内です。偏差を2パーセントポイント未満に保つには、サンプルサイズは20にする必要があります。次に、より歪んだカイ二乗分布 Chi-square(3)と指数分布の結果を検討します。

表8 12個のサンプルを使用したカイ二乗分布と指数分布の Welch 検定のシミュレーション結果

n	カイ二乗分布 Chi-square(3)	指数分布
10	0.1013	0.1064
15	0.0854	0.1079

n	カイ二乗分布 Chi-square (3)	指数分布
20	0.0850	0.0951
30	0.0746	0.0829
40	0.0727	0.0735
50	0.0675	0.0694

これらの大きく歪んだ分布は難しい課題です。目標 $\alpha = 0.05$ からの偏差が3パーセントポイントを優に上回っても許容できるならば、 $n = 15$ は、カイ二乗分布 Chi-square (3)でも十分とみなされる可能性があります。指数分布の場合は $n = 30$ により近いサイズが必要です。特定のサンプルサイズの基準はいくぶん任意であり、 $n = 20$ はさまざまな分布でうまく機能するが、極端に歪んだ分布ではかろうじてうまく機能するに過ぎないという事実は踏まえつつ、サンプル数 10~12 に推奨される最小サンプルサイズとして $n = 20$ を使用します。極端に歪んだ分布でも偏差を小さく保つ必要がある場合は、明らかに、より大きなサンプルが推奨されます。

© 2020 Minitab, LLC. All rights reserved. Minitab®, Minitab Workspace™, Companion by Minitab®, Salford Predictive Modeler®, SPM®, and the Minitab® logo are all registered trademarks of Minitab, LLC, in the United States and other countries. Additional trademarks of Minitab, LLC can be found at www.minitab.com. All other marks referenced remain the property of their respective owners.