

基礎から学ぶシリーズ 4

College Analysis で学ぶ 多変量解析

福井正康

福山平成大学経営学部経営学科

はじめに

このシリーズ、基礎からの統計学では、データの集計方法と検定・推定について少し理論に踏み込んで勉強しました。その際処理はすべて Excel を使い、何を計算しているのか分かるようにしました。ただこの本は経済・経営系の大学院に進もうとする人に基礎を学んでもらう目的で作ったもので、実用を目的とした人向きではありません。

そこで我々は徹底した実用を目的に本を作ることにしました。計算はすべて統計処理のソフトを使い、解説はできるだけやさしく、初心者に応用力をつけることが目的です。数式も定義をはっきりさせるために出てきますが、もちろん飛ばしてもらっても結構です。とにかく「習うより慣れろ」でやってみて下さい。最後には統計処理を見渡せる力がつくことと思います。

さて、統計ソフトにはいろいろなものがあります。SPSS, SAS, S-PLUS, R のように世界的に評価されているものや比較的使い易い STATISTICA, R-Commander 等、数多くのものが開発されています。これらの単独ソフトの他にも Excel の機能を利用するために VBA で記述されたマクロ的なソフトもあります。どれを利用するかは個人の好みでしょうが、一般に上中級者用のものは非常に高価で、初心者用のものでもある程度費用がかかります。またフリーのものでも、R は文系の学生にはちょっと難しいという感じがします。

そこで我々は、学生に自由に使ってもらうために、分かりやすい初心者向けの統計ソフトを開発することにしました。せっかくですからその当時開発中だった OR 関係の分析ソフトに統合させ、できたものが「College Analysis」です。「分析」という大げさな名前ですので、今後より多くの分析手法を加えて充実させていかなければなりません。これはインターネット上で公開していますので、いつでも最新のものを自由に利用することができます。

この教科書では、我々の開発した分析ソフト College Analysis を使って多変量解析を実行する方法を学びます。章建ては基本的に 1 分析 1 章としますが、多変量解析に含まれる分析手法はほぼ独立なので、章ごとにどこからでも読めるようにします。理論は定義をまとめる程度で、使い方と意味を集中して学びます。

福山平成大学 福井正康

多変量解析とは

多変量解析とは複数のデータ間の関係を調べる分析手法の総称です。例えば模擬試験と入試を例に考えてみましょう。まず考えるのが、何回かの模擬試験で入試の点数を予測できないかということです。このように複数の変数（この場合は何回かの模擬試験の結果）で、1つの変数（入試の点数）を予測するような手法を「重回帰分析」といいます。集計と検定のところで述べた回帰分析は、1つの説明変数で目的変数を予測する手法でしたが、この説明変数が複数個になったと考えればよいでしょう。重回帰分析はこの予測式を与える分析手法です。

次に入試の点数までは予測できないでも合否は予測できないものかと考えたとしても、この合否予測をする式を与えるのが「判別分析」です。判別分析は、複数の変数を利用して個体を分類することに利用されます。

模試を受験すると人によって得点に特徴が見られます。成績の良い人、良くない人、理系科目の得意な人、文系科目の得意な人など、様々な特徴があります。これらの特徴を見出す手法が、主成分分析や因子分析です。主成分分析は複数の変数でこの特徴を表す変数（主成分）を作りだす式を与えます。また因子分析は逆にそれぞれの変数はいくつかの特徴的な変数（因子）から影響を受けると考えて、その式を作りだします。主成分や因子は通常変数の数より少なくして、利用者がその意味を考えます。

受験する人を似た者同士分類したり、受験科目を分類したりするにはクラスター分析が用いられます。しかしこれは何を持って似ているのかということの基準が利用者それぞれなので、結果が結構主観的なように思えます。複数の変数同士の類似性は正準相関分析によって与えられます。

これまでは量的データについての解析でしたが、質的なデータについて重回帰分析に相当するものが、数量化Ⅰ類です。また、判別分析に相当するものは数量化Ⅱ類です。質的データについて主成分分析のように分類に使われるものは数量化Ⅲ類と対応分析です。

2000年以降非常によく利用され始めたのが共分散構造分析です。これは変数間の因果関係をモデルとして表す分析手法で、これにより複雑なモデルの解釈ができるようになりました。

多変量解析は複数の変数の関係を与える分析手法の集まりなので、ここで述べるのは適切かどうか分かりませんが、この本では、実験計画法と時系列分析についても解説しようと思います。実験計画法には1元比較（1元配置）実験計画法や2元比較（2元配置）実験計画法（原理的には一般のn元配置もありますが、現実的には2元配置ま

で良く利用されます。) があります。1 元配置は 2 群間の差の検定を複数群間に拡張したものと解釈してもらえればよいと思います。2 元配置は群分けを 2 種類としたようなもので、成績を見るのに性別と勉強量 (分類) で分けて比べるようなものです。1 元配置と違うところは、単純に 2 つの群分けを重ねるのではなく、性別と勉強量の交互作用というものも考えられるようになります。

時系列分析は時間の経過とともに並んだデータから特徴を見出し、未来を予測する手法です。特にデータの変動を特徴的な変動の合成とみて、変動を分解して行く方法がよく利用されます。この本ではこの変動の分解モデルと呼ばれる方法について説明します。

以上たくさんの分析がありますが、最初に集計と検定の続きとして実験計画法から始めましょう。

1 章 実験計画法

College Analysis のメニュー [分析-基本統計-量的データの検定-量的データの検定メニュー] と [-質的データの検定-質的データ検定メニュー] を選んで表示される図 1.1 画面の中で、赤枠で示された部分の違いを見て下さい。

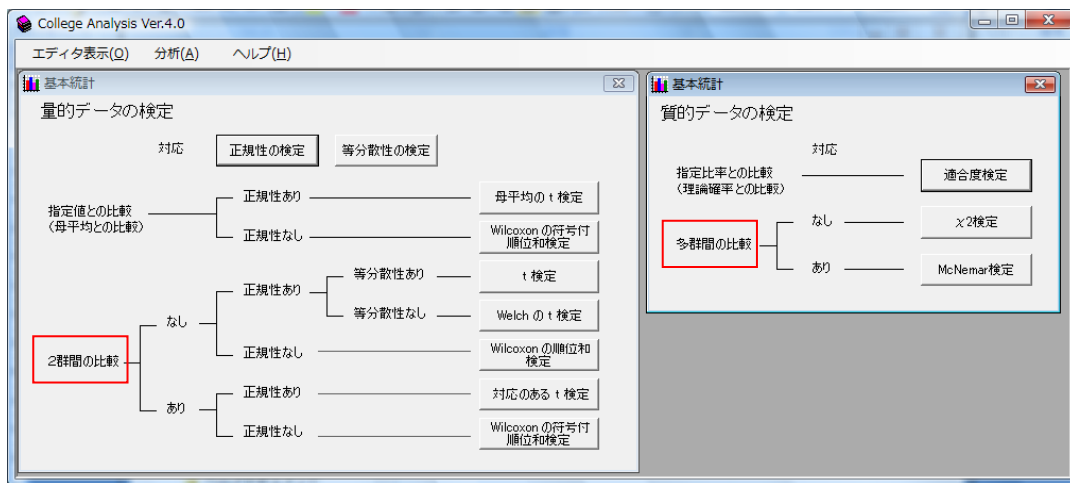


図 1.1 量的データと質的データの比較

質的データについては多群間となっていますが、量的データについては 2 群間になっています。ここでの検定手法は、質的データについては一般に複数の群の間で利用できますが、量的データについては 2 群間に限定されていたのです。では量的データで多群間の比較を行うにはどうすればよいのでしょうか。この答えがここで述べる実験

計画法です。

1.1 1元配置実験計画法

多群間の平均や中間値に差があるかどうか検討する手法であり、変数を比較する属性の数によって1元配置（1元比較ともいう）、2元配置（2元比較ともいう）などと分かれています。特に2元配置以上では属性間の交互作用による影響も考察しますが、ここでは理解し易い1元配置についてのみ解説するに留めます。

検定は2群間の比較と同じように、正規性や等分散性の有無によって2つの手法に分れます。それを表したものが図1.1です。

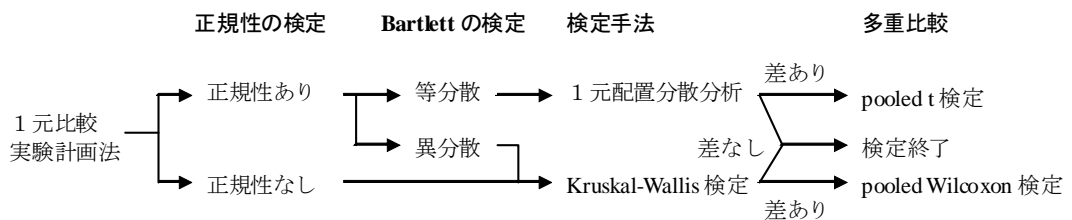


図 1.1.1 1元比較実験計画法の構造

まず最初に各群のデータが正規性をするかどうか検定します。次に、すべての群で正規性が認められたとき、等分散性の検定に移ります。等分散性の検定では、これまでの2群間比較のF検定ではなく、多群間で分散を比較するBartlettの検定という手法が利用されます。これで等分散であると認められた場合（実際には異分散といえないと判定された場合）には、1元配置分散分析と呼ばれる手法が利用されます。正規性が認められなかったり、認められた場合でも異分散であると判定された場合は、データの順位を利用したノンパラメトリックな手法であるKruskal-Wallisの検定を行います。

実際のCollege Analysisの画面は図1.1.2になります。まず最初に取り組むのは赤枠で囲まれた部分です。

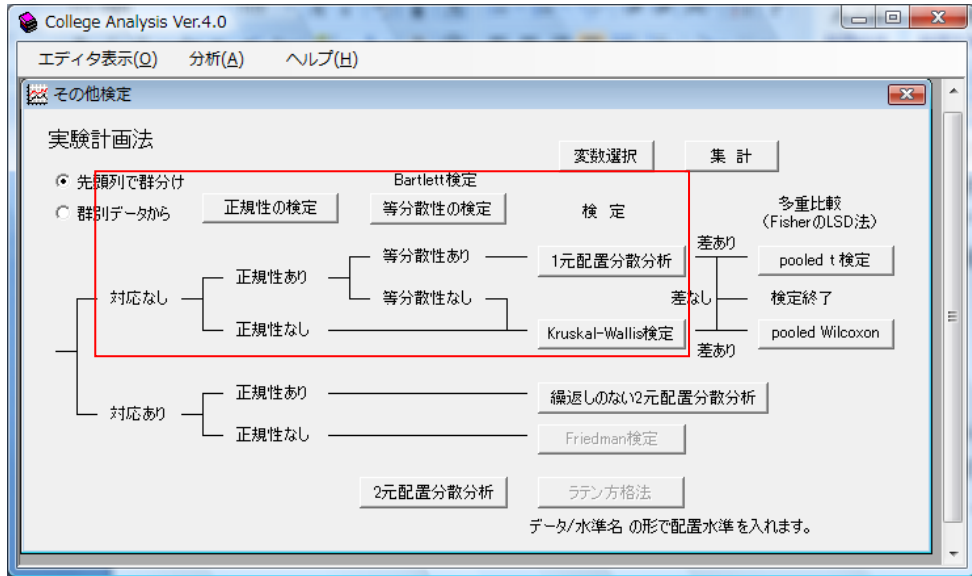


図 1.1.2 実験計画法分析メニュー

1 元配置分散分析や Kruskal-Wallis 検定で有意差が出た場合、どの要素間に差があるのか興味湧きます。しかしこれには少し難しい問題があり、後の多重比較の節までお待ち下さい。ここではまず 1 元配置分散分析から見て行くことにしましょう。

1.2 1 元配置分散分析

ここではまず 1 元配置分散分析について見てみましょう。これは各群のデータに正規性があり、等分散である場合にのみ利用できる最も差を見つけ易い検定手法です。以下の例を見てみましょう。

例

3つの条件である商品の売上を調査したところ、以下の結果を得た。(Samples¥分散分析 ex.txt) これらの分布が正規分布で条件間で等分散であることを仮定して、条件間に差があるといえるか、有意水準 5% で判定せよ。

条件 1 115, 110, 108, 114, 120, 116, 108, 112, 115, 122
 条件 2 121, 118, 124, 117, 119, 130, 121, 115, 118, 119
 条件 3 116, 112, 120, 111, 112, 108, 114, 119, 104, 113

解答

データを読み込み、「群別データ」からのラジオボタンをチェックし、変数選択ですべ

て選択して、「1元配置分散分析」ボタンをクリックすると、図 1.2.1 のような結果が表示されます。

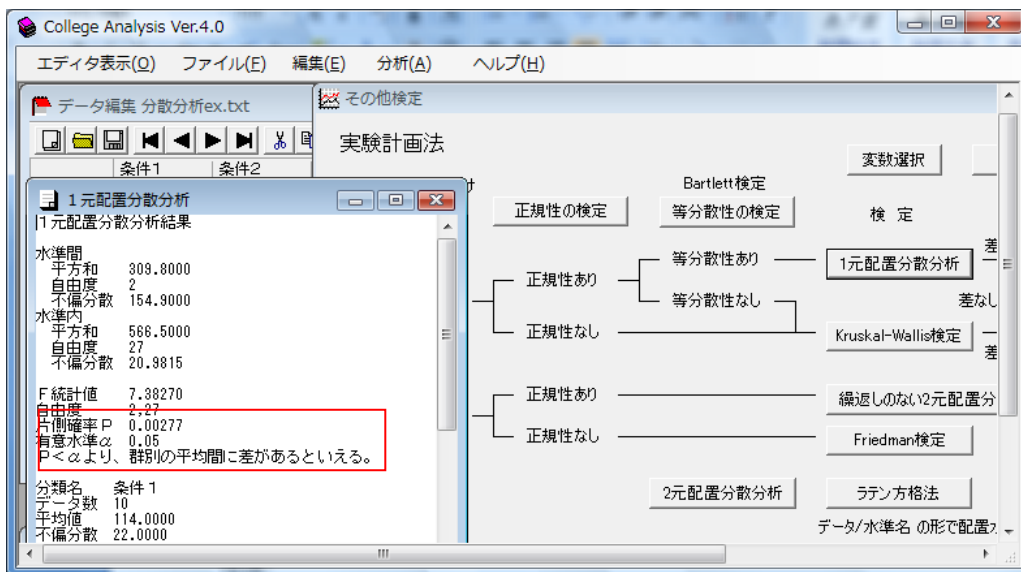


図 1.2.1 1元配置分散分析結果

これは、3群間に差があるかどうかを示すもので、赤枠の部分に注目します。これと同時にテキスト出力の下に、グリッドで図 1.2.2 のような結果も出力されます。

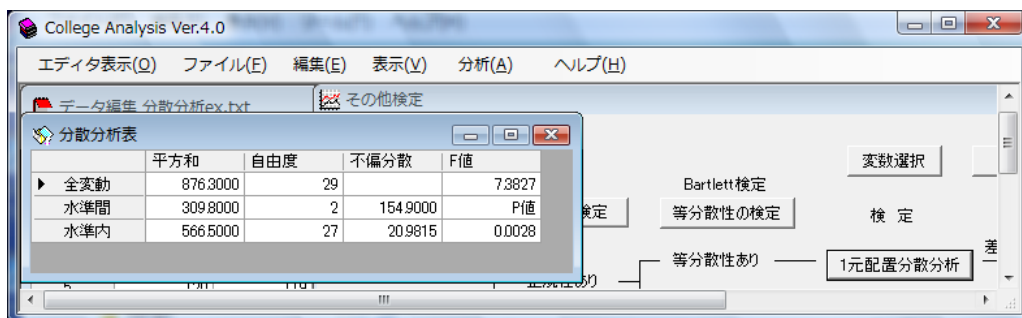


図 1.2.2 分散分析表

これは分散分析表と呼ばれるもので、結果はテキスト出力のものと同じです。ここでは以下のような理論を使っています。

理論

水準間に差があるかどうか、有意水準 α で検定する。

水準 1	水準 2	...	水準 k
x_{11}	x_{21}	...	x_{k1}
x_{12}	x_{22}	...	x_{k2}
\vdots	\vdots	\vdots	\vdots
x_{1n_1}	x_{2n_2}	...	x_{kn_k}

$x_{i\lambda}$ は水準 i に固有な値 μ_i と誤差 $\varepsilon_{i\lambda}$ とからなると仮定する。

$$x_{i\lambda} = \mu_i + \varepsilon_{i\lambda} \quad \varepsilon_{i\lambda} \sim N(0, \sigma^2) \text{ 分布}$$

全変動は以下のように分解される。

$$S = \sum_{i=1}^k \sum_{\lambda=1}^{n_i} (x_{i\lambda} - \bar{x})^2 = \sum_{i=1}^k \sum_{\lambda=1}^{n_i} (x_{i\lambda} - \bar{x}_i)^2 + \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 = S_E + S_P$$

全変動 水準内変動 水準間変動

そのとき、各変動の分布は以下となる。

$$S/\sigma^2 \sim \chi_{N-1}^2 \text{ 分布}, \quad S_E/\sigma^2 \sim \chi_{N-k}^2 \text{ 分布}, \quad S_P/\sigma^2 \sim \chi_{k-1}^2 \text{ 分布}, \quad \text{ここに } N = \sum_{i=1}^k n_i$$

帰無仮説 $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ (水準間に差がない)

対立仮説 $H_1: H_0$ でない

帰無仮説のもとで

$$F = \frac{S_P/(k-1)}{S_E/(N-k)} \sim F_{k-1, N-k} \text{ 分布}$$

$F = F_{k-1, N-k}(p)$ として、 $p < \alpha$ ならば、水準間に差があると判定する。

1.3 Kruskal-Wallis 検定

次は正規性がなかった場合や、正規性があっても等分散でなかった場合に利用する Kruskal-Wallis 検定です。この検定は分布形によらない検定ですので、正規性と等分散性が成り立つ場合に使用しても問題はありません。ただ、結果は 1 元配置分散分析の方が良いというだけです。ここでは 1.2 節と同じデータを使います。

例

3つの条件である商品の売上を調査したところ、以下の結果を得た (1.2 節参照)。分布が正規分布に従わないとして、これらの条件間に差があるかどうか有意水準 5% で

判定せよ。

解答

図 1.1.2 の分析メニューで「Kruskal-Wallis 検定」ボタンをクリックすると、図 1.3.1 のような結果が表示されます。

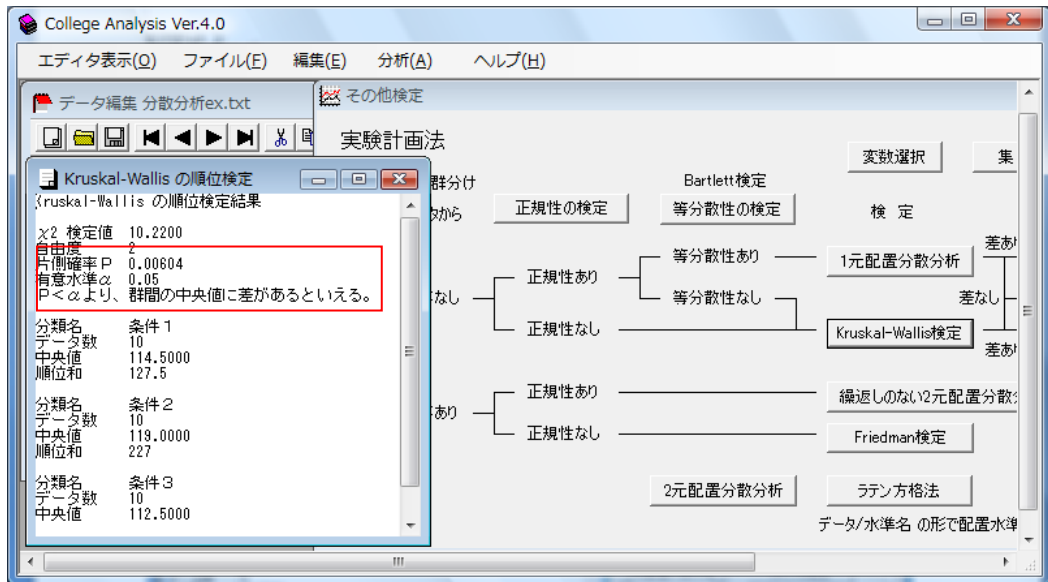


図 1.3.1 Kruskal-Wallis 検定結果

この分析で使う理論は以下の通りです。

理論

k 種類の水準の中間値に差があるかどうか、有意水準 $\alpha \times 100\%$ で判定する。
全データの小さい順に順位を付ける。

水準 1	水準 2	...	水準 k
r_{11}	R_{21}	...	r_{k1}
r_{12}	R_{22}	...	r_{k2}
\vdots	\vdots	...	\vdots
r_{1n_1}	r_{2n_2}	...	r_{kn_k}
w_1	w_2	...	w_k

水準毎のデータ数 n_i , $N = \sum_{i=1}^k n_i$, 水準毎の合計 w_i

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k n_i \left(\frac{w_i}{n_i} - \frac{N+1}{2} \right)^2 \sim \chi_{k-1}^2 \text{ 分布}$$

$\chi^2 = \chi_{k-1}^2(p)$ として、 $p < \alpha$ ならば、水準間に差があると判定する。

1.4 等分散性の検定 (Bartlett 検定)

前節までは分布を仮定して検定を行いましたが、ここでは分布を決めることを考えます。まず、「正規性の検定」ボタンをクリックして、量的データの検定メニューを表示し、正規性の検定を行います。群分けするデータか、元々群別のデータかを十分確認して下さい。

すべての群で正規性が認められたら（正確には非正規でなかったら）正規性があると判断します。1つでも正規分布といえないと判断されたら、正規分布でないと判断します。正規性の検定は正規分布であると積極的に言えませんので、ヒストグラムや正規確率紙（Q-Q プロット）の方法も併用します。

正規性が認められた場合は、次に等分散性の検定を行います。「等分散性の検定」ボタンをクリックすると結果が表示されます。それでは以下の例をやってみましょう。

例

3つの条件である商品の売上を調査したところ、1.2節の結果を得た。1元配置分散分析と Kruskal-Wallis 検定のどちらを利用するか判断せよ。

解答

まずこのデータについて正規性の検定をしてみましょう。群別データであることを確認して実施すると3つの群で正規性ありとみなすの判定です。そこで等分散性の検定 (Bartlett の検定) を行くと、以下の結果になります。

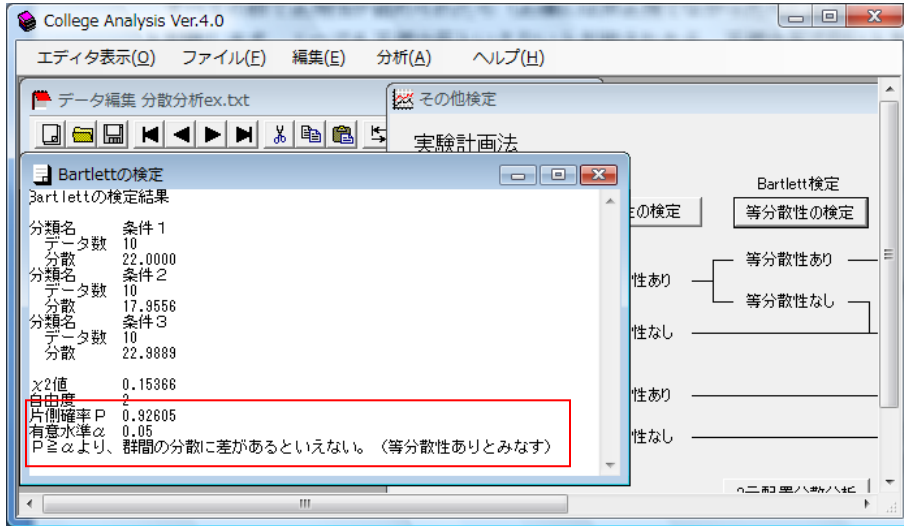


図 1.4.1 Bartlett の検定結果

この結果から等分散性があると判断します。検定は等分散であると断言しているわけではないので注意して下さい。ここで使った Bartlett の検定の理論は以下になります。

理論

帰無仮説 $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$

対立仮説 $H_1 : H_0$ でない

$$V_E = \frac{S_E}{n-k} = \frac{1}{n-k} \sum_{i=1}^k \sum_{\lambda=1}^{n_i} (x_{i\lambda} - \bar{x}_i)^2, \quad S_i^2 = \frac{1}{n_i - 1} \sum_{\lambda=1}^{n_i} (x_{i\lambda} - \bar{x}_i)^2$$

$$C = 1 + \frac{1}{3(k-1)} \left[\sum_{j=1}^k \frac{1}{n_j - 1} - \frac{1}{n-k} \right] \quad \text{とすると、}$$

$$\chi^2 = \frac{1}{C} \left[(N-k) \log V_E - \sum_{i=1}^k (n_i - 1) \log S_i^2 \right] \sim \chi_{k-1}^2 \text{ 分布}$$

$\chi^2 = \chi_{k-1}^2(p)$ として、 $p < \alpha$ ならば、水準間に差があると判定する。

問題 1

Samples¥分散分析 1.txt は 3 つの工場群の不良品率を与えたものである。各群に差があるといえるか、実験計画法を用いて有意水準 5% で検討せよ。

正規性の検定 正規分布と [みなす・いけない]

等分散性の検定 検定確率 [] 等分散と [みなす・いけない]

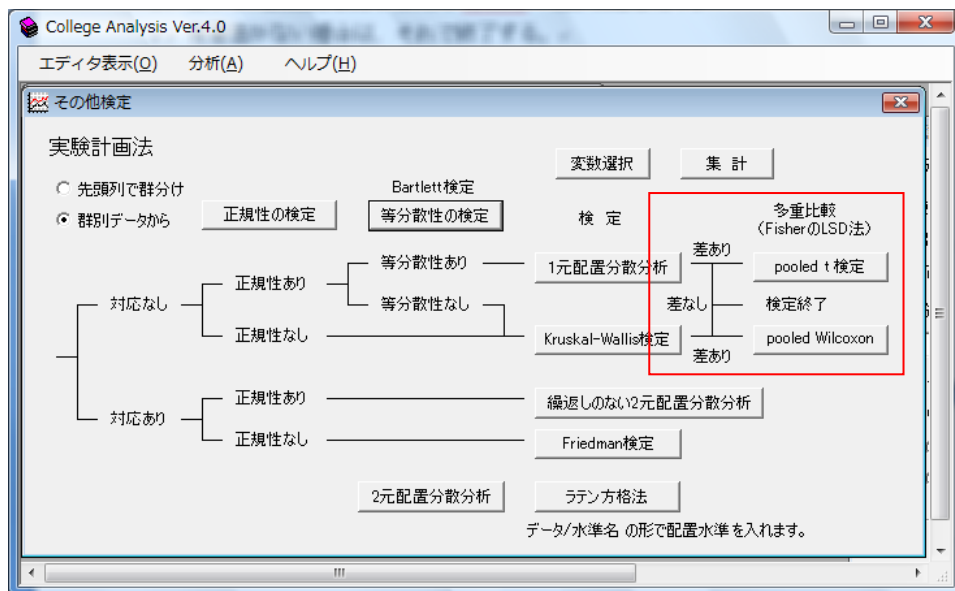


図 1.5.1 多重比較部分

1.5.1 正規性・等分散性のある場合の多重比較

例

3つの条件である商品の売上を調査したところ、1.2節の結果を得た。(Samples¥分散分析 ex.txt) これらの分布が等分散の正規分布であるとして、分散分析によって条件間に差があると判定された。ではどの条件間に差が見られるのだろうか、有意水準5%で判定せよ。

解答

正規性・等分散性が認められる場合ですから、pooled 推定値を用いた t 検定を利用します。これは通常の2群比較を行う分散を、すべての群から見積もられた分散(pooled 推定値)に置き換える手法です。図 1.1.2の「pooled t 検定」ボタンをクリックすると以下の結果が表示されます。

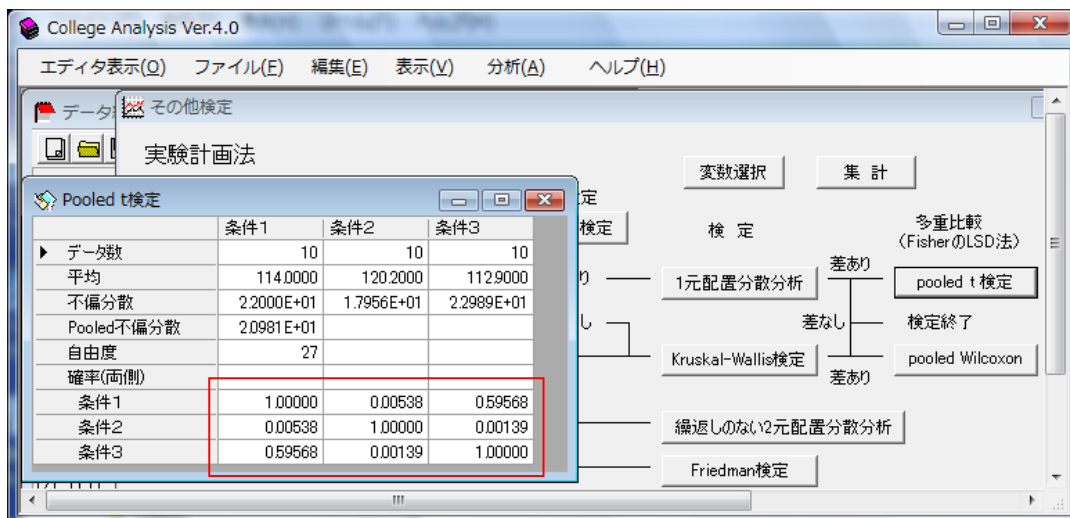


図 1.5.1 pooled 推定値を用いた t 検定結果

下の赤枠の部分が検定確率で、左と上の変数名の部分で読みます。例えば、条件 1 と条件 2 の差の検定確率は 0.00538 になっています。この結果から、条件 1 と条件 2、条件 2 と条件 3 の間に有意差があると判断します。ここで用いた理論は以下の通りです。

理論 (pooled 推定値を用いた t 検定)

k 種類の水準を考え、各水準の平均の間に差があるか有意水準 $\alpha \times 100\%$ で判定する。水準 i のデータ数を n_i 、平均を \bar{x}_i 、不偏分散を u_i^2 として、水準 i, j について考える。

$$N = n_1 + n_2 + \dots + n_k$$

$$u^2 = \frac{(n_1 - 1)u_1^2 + (n_2 - 1)u_2^2 + \dots + (n_k - 1)u_k^2}{N - k} \quad \text{pooled 不偏分散}$$

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{u \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{N-k} \text{ 分布} \quad (\text{t 検定統計量の不偏分散についての拡張})$$

$t_{ij} = t_{N-k}(p/2)$ として $p < \alpha$ ならば、水準間に差があると判定する。

1.5.2 正規性のない場合の多重比較

次は、正規性がない場合と正規性があっても等分散でない場合の多重比較法である結合順位を用いた Wilcoxon 順位和検定 (College Analysis の中では pooled Wilcoxon と表されています) について説明します。以下の例を見て下さい。

例

3つの条件である商品の売上を調査したところ、1.2 節の結果を得た。(Samples分散分析 ex.txt) これらの分布は正規分布でないとして、Kruskal-Wallis 検定によって条件間に差があると判定された。ではどの条件間に差が見られるのだろうか、有意水準 5% で判定せよ。

解答

図 1.1.2 の分析メニューで「pooled Wilcoxon」ボタンをクリックすると以下の結果が表示されます。

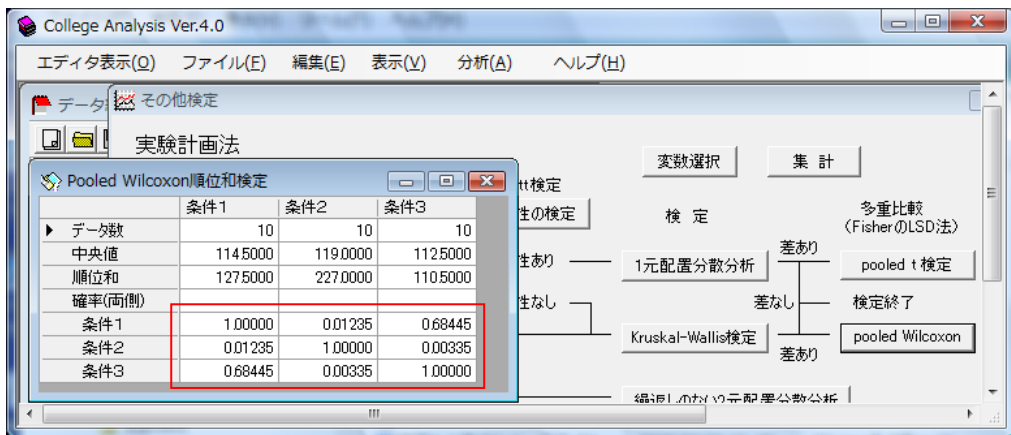


図 1.5.2 結合順位を用いた Wilcoxon の順位和検定

下の赤枠の部分が検定確率です。これによると条件 1 と条件 2、条件 2 と条件 3 の間に有意差が見られます。ここで用いた理論は以下の通りです。

理論（結合順位による Wilcoxon の順位和検定）

k 種類の水準のどの中間値に差があるか、有意水準 $\alpha \times 100\%$ で判定する。
全データの小さい順に順位を付ける。

水準 1	水準 2	...	水準 k
r_{11}	r_{21}	...	r_{k1}
r_{12}	r_{22}	...	r_{k2}
\vdots	\vdots	...	\vdots
r_{1n_1}	r_{2n_2}	...	r_{kn_k}

w_1	w_2	...	w_k
-------	-------	-----	-------

水準毎のデータ数 n_i , $N = \sum_{i=1}^k n_i$, 水準毎の合計 w_i データ数は十分多いとする。

$$Z_{ij} = \frac{\left| \frac{w_i}{n_i} - \frac{w_j}{n_j} \right| - \frac{1}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}{\sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim N(0,1) \text{ 分布}$$

$Z_{ij} = Z(p/2)$ として、 $p < \alpha$ ならば、水準間に差があると判定する。

問題 3

Sample¥分散分析 1.txt は 3 つの工場群の不良品率を与えたものである。各群に差があるといえるか、実験計画法を用いて有意水準 5% で検討せよ。

正規性の検定 正規分布と [みなす・いえない]

等分散性の検定 検定確率 [] 等分散と [みなす・いえない]

検定名 [] 検定確率 []

判定 工場群間の不良品率に差があると [いえる・いえない]

差があるとするとの条件間に差があるか。差がある条件同士を工場 2 < 工場 3 (これは実際の結果とは関係ない) のように不等号で表せ。

検定名 []

結果 []

問題 4

Sample¥分散分析 2.txt は 4 つの群のデータであるが、各群に差があるといえるか、実験計画法を用いて有意水準 5% で検討せよ。

正規性の検定 正規分布と [みなす・いえない]

等分散性の検定 検定確率 [] 等分散と [みなす・いえない]

検定名 [] 検定確率 []

判定 群間に差があると [いえる・いえない]

差があるとするとの群間に差があるか。差がある群同士を群 2 < 群 3 (これは実際の結果とは関係ない) のように不等号で表せ。

検定名 []

結果 []

はそれぞれ repeated measured 1元配置分散分析、repeated measured Kruskal-Wallis 検定とも呼ばれています。実際に例を見てみましょう。

例

3つの条件である商品の売上を調査したところ、1.2節の結果を得た。各データに対応があるとして差があるか検定せよ。(再掲)

条件1 115, 110, 108, 114, 120, 116, 108, 112, 115, 122

条件2 121, 118, 124, 117, 119, 130, 121, 115, 118, 119

条件3 116, 112, 120, 111, 112, 108, 114, 119, 104, 113

解答

ファイル Samples¥分散分析 ex.txt を読み込んだ後、変数選択をして正規性の有無を調べるために、「正規性検定」ボタンをクリックします。表示された量的データの集計メニューで「対応のあるデータから」ラジオボタンを選択し、「S-W 検定*」のボタンをクリックすると図 1.6.2 の結果が表示されます。

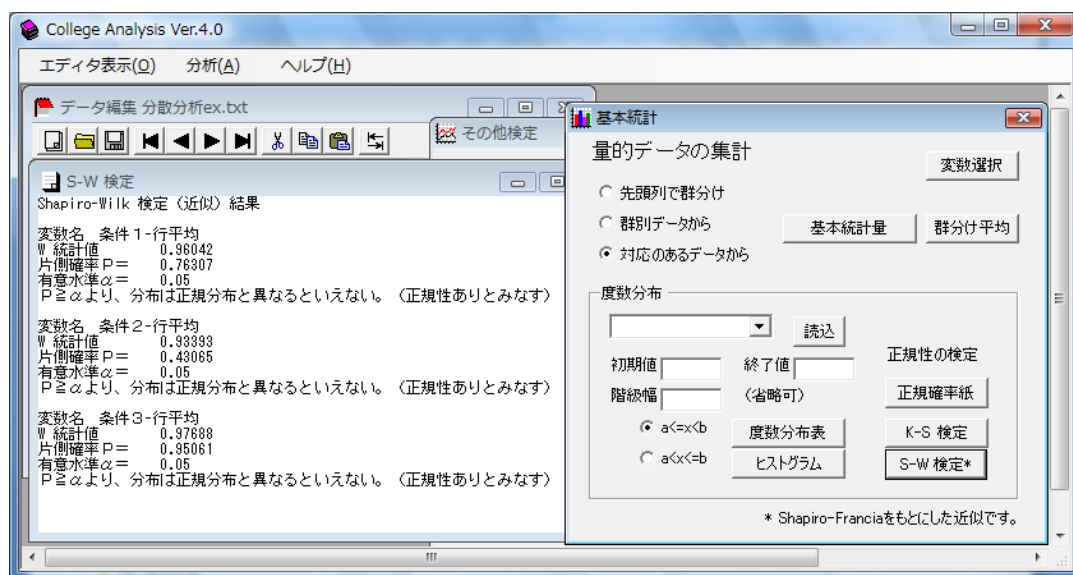


図 1.6.2 正規性の検定結果

ここでは各変数から行の平均を引くという処理を行って正規性の判定をしています。2変数の場合は、差を取る処理に相当します。「正規性ありとみなす」という判定なの

で、「繰り返しのない2次元配置分散分析」ボタンをクリックします。結果は図 1.6.3 のようになります。

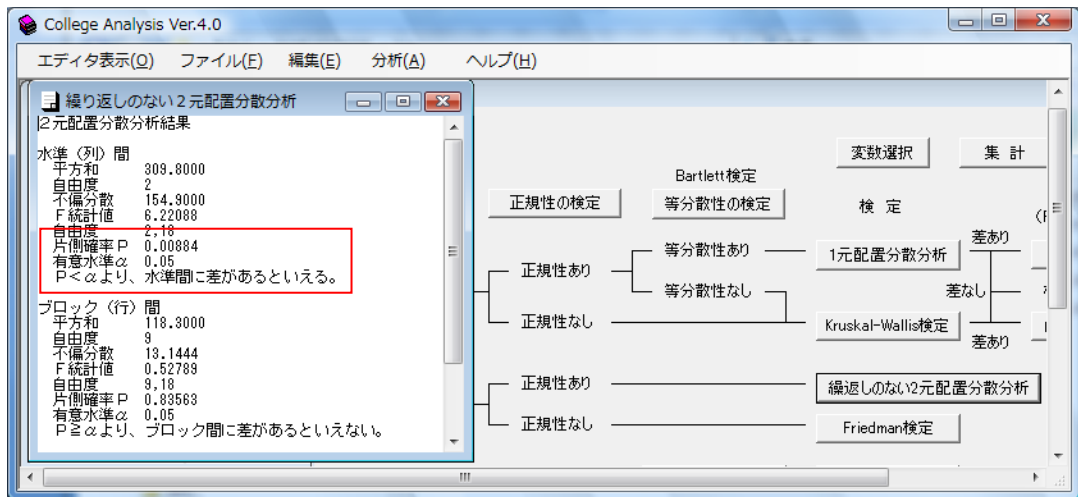


図 1.6.3 繰り返しのない2次元配置分散分析結果

赤枠で囲まれた部分に注目して下さい。これより条件間に差があるといえるという判定になります。

もし、正規性の検定で正規分布と言えないと判定されたら、「Friedman 検定」ボタンをクリックします。すると図 1.6.4 のように表示されます。

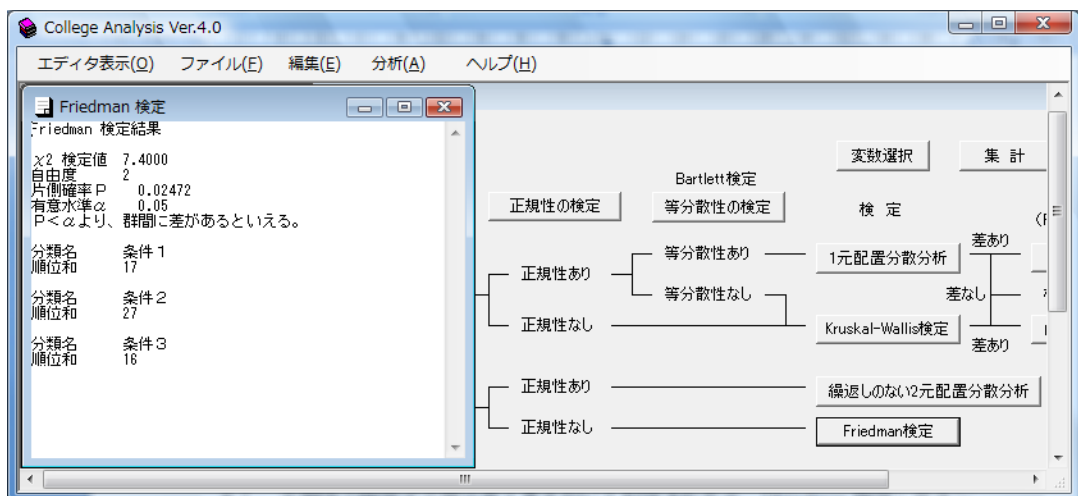


図 1.6.4 Friedman 検定結果