

統計ソフトウェア R と HAD の使用方法

本資料は、多田実・多田伶（2025）『事例で学ぶマーケティング・リサーチ：定性・定量調査の理論と実践』 オーム社におけるソフトウェアの使い方を解説するものです。

ここでは、統計ソフトウェア R¹、表計算ソフトウェア Excel 用の無料ツール HAD²を用いて、差の分析（第 9 章）、回帰分析（第 10 章）、因子分析（第 11 章）、クラスター分析（第 12 章）の実行方法を示します。掲載する分析例はすべて **Windows 環境**で行ったものであり、**R のバージョンは 4.4.0、HAD のバージョンは 18_010** です。ソフトウェアの基本的な使い方については、序章および第 9 章～第 12 章に掲載した文献案内の情報も確認してください。

第 9 章 差の分析

第 9 章では、「世界価値観調査」の統計データ（9_anova.csv）を利用して、分散分析を行います。CSV ファイルの一部を以下に掲載しておきますので、確認してください。

分析データ（一部抜粋）

	A	B	C	D
1	番号	国籍	世代	信用度
2	1	Japan	A4	3
3	2	Japan	A4	3
4	3	Japan	A4	3
5	4	Japan	A1	3
6	5	Japan	A3	1
7	6	Japan	A4	3
8	7	Japan	A2	2
9	8	Japan	A1	2
10	9	Japan	A3	2
11	10	Japan	A1	3

(1) R の使い方

R でのデータの読み込み方はいくつかありますが、ここでは、CSV ファイルを直接 R に入れるやり方を紹介します。筆者はパソコンのデスクトップ上に「Support」というファイルを作り、データを保管しました。各自でファイルを作成し、データを管理してください。

データを読み込む際、**ディレクトリ**（パソコンにおける住所のようなもの）を指定する必要があります。「ファイル」内にある「ディレクトリの変更」をクリックしましょう。今、デスクトップ上の「Support」にデータを保管していますので、それらを選び、「フォルダー

¹ R. <https://cran.r-project.org/bin/windows/base/>（2025 年 10 月 31 日閲覧）

² 統計分析ソフト HAD. <https://norimune.net/had>（2025 年 10 月 31 日閲覧）

の選択」をクリックします。setwd 関数でディレクトリを指定しても問題ありません。

次に、read.csv 関数でデータを読み込みますが、最初にファイル名を入力します。「header = TRUE」で先頭行が変数名（ラベル）であるという指示を与え、「fileEncoding = "CP932"」で文字コードを指定しています。Windows 環境で日本語を含むデータを読み込む際、Shift-JIS (CP932) を指定すれば、エラーは解消されるはずです。head 関数を用いると、デフォルトで上から 6 行が出力されますので、データを入力できたかを確認してみてください。

R でのデータ読み込み

```
> data<- read.csv("9_anova.csv", header=TRUE, fileEncoding="CP932")
> head(data)
  番号  国籍 年代 信用度
1    1   Japan  A4      3
2    2   Japan  A4      3
3    3   Japan  A4      3
4    4   Japan  A1      3
5    5   Japan  A3      1
6    6   Japan  A4      3
```

では、**分散分析**を行ってみましょう。データは 2 要因で、国籍は 2 水準（日本 vs 韓国）、年代は 4 水準（1 層 vs 2 層 vs 3 層 vs 4 層）で構成されていますので、**二元配置分散分析**を適用します。やり方はいくつかありますが、aov 関数を用いたときの結果は以下の通りです³。aov 関数における「data=」の後ろにはデータ名を入力します。今回、data という名称で、CSV ファイルを読み込んでいますので、「data = data」となります。

「~」の左側には**目的変数**（テレビの信用度）、右側には**説明変数**（国籍と年代）を入力します。説明変数が 2 つ以上ある場合は「+」でつなげ、変数間の**交互作用効果**は「:」で表記します。「*」を用いると、主効果と交互作用効果を同時に考慮できますので、「国籍*年代」で国籍の主効果、年代の主効果、国籍と年代の交互作用効果を検討していることになります。

p 値は「 $P_r(> F)$ 」のところに記載されています。p 値を見ると、国籍と年代の主効果は 0.001%水準で有意、国籍と年代の交互作用効果は非有意であることがわかります。これより、韓国人よりも日本人のテレビの信用度が高いことが示されました。

³ aov 関数では、タイプ I 平方和が出力されます。car パッケージの Anova 関数を用いると、タイプ I~III を選択することができます。

R における分散分析の結果

```
> # 二元配置分散分析
> res<- aov(信用度~国籍*年代,data=data)
> summary(res)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
国籍	1	6.9	6.900	17.732	2.63e-05 ***
年代	3	15.7	5.243	13.475	1.01e-08 ***
国籍:年代	3	1.5	0.503	1.294	0.275
Residuals	2505	974.8	0.389		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

年代は 4 水準で構成されており、どの年代間に差があるかは不明ですので、**チューキ・クレイマー法の多重比較検定**を行います。TukeyHSD 関数に先ほどの aov 関数を投入すれば、多重比較検定を実行できます。**p 値**は「p adj」に示されており、diff は水準間の差を表しています。分析の結果、2 層と 3 層以外で、年代間に差があることが示されました。

R における多重比較検定の結果

```
> # 多重比較検定
> TukeyHSD(aov(信用度~年代,data=data))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = 信用度 ~ 年代, data = data)

$年代
```

	diff	lwr	upr	p adj
A2-A1	0.09758307	0.003509529	0.1916566	0.0385835
A3-A1	0.11777362	0.024474410	0.2110728	0.0065278
A4-A1	0.27478238	0.175632345	0.3739324	0.0000000
A3-A2	0.02019055	-0.063839163	0.1042203	0.9264623
A4-A2	0.17719931	0.086717381	0.2676812	0.0000031
A4-A3	0.15700877	0.067332173	0.2466854	0.0000418

(2) HAD の使い方

HAD に「世界価値観調査」の統計データを読み込んでみましょう。HAD を立ち上げて、セキュリティリスクのエラーメッセージ（このファイルのソースが信頼できないため、Microsoft によりマクロの実行がブロックされました。）が出た場合、以下の手続きを行ってください。HAD のファイルを右クリックし、「プロパティ」を選択します。プロパティ内の「セキュリティ」の「許可する」にチェックを入れ、OK を押してください。その後、HAD を立ち上げると、エラーメッセージは出現せず、ソフトを利用できるようになります。

HAD には 2 つのシートがあり、1 つ目が「データ」画面、2 つ目が「モデリング」画面です。データを読み込む際、1 つ目のシートを使います。以下のようにデータを入力できたならば、「変数名」の下に「データ読み込み」をクリックしてください。

HAD のプロパティ画面



HAD のデータ読み込み画面

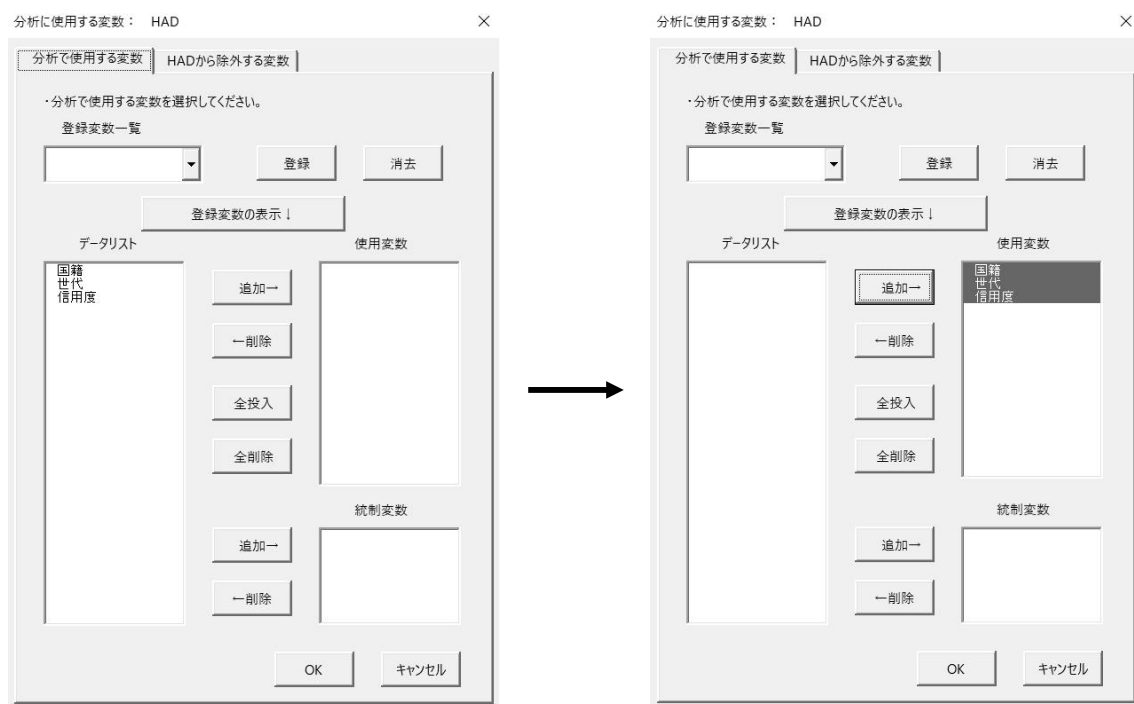
	A	B	C	D	E
1	変数名	番号	国籍	世代	信用度
2	データ読み込み	1	Japan	A4	3
3		2	Japan	A4	3
4		3	Japan	A4	3
5		4	Japan	A1	3
6	モデリングシート	5	Japan	A3	1
7		6	Japan	A4	3
8		7	Japan	A2	2
9		8	Japan	A1	2
10	列幅の調整	9	Japan	A3	2
11		10	Japan	A1	3
12		11	Japan	A2	2
13		12	Japan	A3	3
14		13	Japan	A4	3

データの読み込み後、モデリング画面の「変数情報」に番号、国籍、世代、信用度が入っていることを確認しましょう。次に、「使用変数」をクリックして、「データリスト」から国籍、世代、信用度を選択します。「追加→」を押すと、「使用変数」に3つの変数名が入ります。これらを確認できたならば、OKを押して、次に進みましょう。モデリング画面の「使用変数」が番号、国籍、世代、信用度となっているかについてもチェックしてください。

HAD のモデリング画面

	A	B	C	D	E	F	G	H	I	J
1	変数名	番号	国籍	世代	信用度					
2										
3	分析		選択セルを使用		統制変数を投入		変数を左につめる		<input checked="" type="radio"/> 変数情報	
4									<input type="radio"/> 回帰分析	
5									<input type="radio"/> 因子分析	
6	データ									
7	シート									
8										
9	使用変数	番号								
10										
300	変数情報	フィルタ	値ラベル	ラベル	@変数	コード				
301	番号									
302	国籍									
303	世代									
304	信用度									

HAD における変数の設定



モデリング画面の「回帰分析」を選択した後、「モデル保存」横の「分散分析」を選択しましょう。分析の目的は、テレビの信用度が国籍と年代で異なるかを検証することです。データは2要因で、国籍は2水準（日本 vs 韓国）、年代は4水準（1層 vs 2層 vs 3層 vs 4層）で構成されていますので、二元配置分散分析を実行します。

「使用変数」横の「信用度」にカーソルを合わせ、「目的変数を投入」をクリックします。

また、国籍と世代にカーソルを合わせ、「主効果を全投入」「交互作用を全投入」をクリックしてください。「目的変数」と「モデル」に以下の情報が入力されているかを確認しましょう。これで準備は終わりましたので、「分析実行」を押して、結果を見ていきます。

HAD の分析実行画面

	A	B	C	D	E	F	G	H	I	J	K	L
1	変数名	番号	国籍	世代	信用度							
2												
3	分析		選択セルを使用		統制変数を投入		変数を左につめる			○ 変数情報		変数の作成
4										● 回帰分析		
5										○ 因子分析		
6	データシート											データセット
7												
8												
9	使用変数	番号	国籍	世代	信用度							
10												
11												
12	分散分析		目的変数を投入		主効果を全投入		交互作用を全投入			オプション		分析実行
13												
14												
15	目的変数→	信用度	\$									
16												
17	モデル→	国籍	世代	国籍*世代								
18												
19	共変量→											
20												
21	反復測定→											
22												
23	スライス→				スライス		各セルの平均値			☐ 共変量を中心化	☐ 一般化効果量	
24										☐ 名義有意水準を表示	☐ 不偏推定効果量	
25												
26	モデル保存	○ 回帰分析	● 分散分析	○ 一般化線形モデル	○ 一般化線形混合モデル							
27												

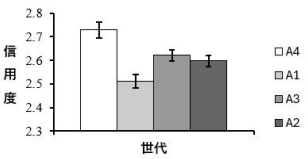
分析を実行すると、新しいシート「Anova」が出現します。「要因の効果（タイプⅢ&平方和）」のところで、国籍と世代の**主効果**、それらの**交互作用効果**の有無がわかります。**p値**を見てみると、主効果はいずれも 0.01%水準で有意、交互作用効果は非有意であることが示されました。国籍については、日本と韓国の2ヵ国間で比較を行っていますので、韓国人よりも日本人のテレビに対する信用度が高いという結果が明らかになりました。

年代については、4水準で構成されているため、どの年代間に差があるかは分散分析だけではわかりません。同画面下の「**多重比較**」のところを確認してみましょう。HADでは、Holm法による**多重比較検定**の結果が出力されます。調整**p値**を見てみると、3層（A3）と2層（A2）に差はありませんが、それ以外の世代間でテレビに対する信用度が異なることが明らかになりました。通常、nsはnot significanceの略語で非有意であることを示します。

HAD における分散分析と多重比較検定の結果

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2	分散分析											分析コード:		0.44秒
3														
4	モデル	信用度	<	国籍	世代	国籍*世代								
5														
6	モデル適合													
7														
8			SS	df	MS	F 値	p 値							
9	モデル	24.140	7	3.449	8.862	.000								
10	誤差	974.751	2505	0.389										
11	全体	998.891	2512											
12														
13														
14		R	R ²	Adjust R ²	AIC	AICC	BIC							
15	適合指標	.155	.024	.021	4769.65	4769.70	4814.28							
16														
17														
18	要因の効果(タイプ III & 平方和)													
19														
20	変数名	SS	MS	MSe	偏 η^2	95%CI	F 値	df1	df2	p 値				
21	国籍	2.614	2.614	0.389	.003	.000, .008	6.717	1	2505	.010 **				
22	世代	9.636	3.212	0.389	.010	.003, .018	8.254	3	2505	.000 **				
23	国籍*世代	1.510	0.503	0.389	.002	.000, .005	1.294	3	2505	.275				

53	要因:世代													
54														
55	水準ごとの平均値													
56	水準	平均値	標準誤差	95%下限	95%上限	t 値	df	p 値						
57	A4	2.729	0.034	2.664	2.795	81.446	2505	.000						
58	A1	2.512	0.029	2.455	2.568	86.904	2505	.000						
59	A3	2.622	0.023	2.576	2.667	113.654	2505	.000						
60	A2	2.599	0.023	2.553	2.645	111.037	2505	.000						
61														
62														
63	多重比較	Holm法						主効果p 値	.000 **					
64	水準の組	差	標準誤差	95%下限	95%上限	t 値	df	p 値	調整p 値					
65	A4 - A1	0.218	0.044	0.131	0.305	4.922	2505	.000	.000 **					
66	A4 - A3	0.108	0.041	0.028	0.187	2.646	2505	.008	.025 *			※エラーバーは標準誤差		
67	A4 - A2	0.131	0.041	0.051	0.211	3.197	2505	.001	.007 **					
68	A1 - A3	-0.110	0.037	-0.183	-0.038	-2.979	2505	.003	.012 *					
69	A1 - A2	-0.087	0.037	-0.160	-0.014	-2.342	2505	.019	.038 *					
70	A3 - A2	0.023	0.033	-0.041	0.087	0.701	2505	.483	ns					



第 10 章 回帰分析

第 10 章では、日本プロ野球の統計データ (10_regression_kyojin.csv) を用いて、回帰分析を行います。CSV ファイルの一部を以下に掲載しておきますので、確認してください。

以下、巨人のデータを利用して、解説を行います。同じように、阪神のデータ (10_regression_hanshin.csv) でも回帰分析を実行できます。各自で試してみましょう。

分析データ (一部抜粋)

	A	B	C	D	E	F	G	H	I
1	番号	入場率	休日	伝統の一戦	交流戦	貯金	応援グッズ	野球グッズ	プレゼント
2	1	0.300	0	0	0	5	0	0	
3	2	0.309	0	0	0	6	0	1	
4	3	0.462	0	0	0	2	0	0	
5	4	0.484	0	0	0	5	0	1	
6	5	0.488	0	0	0	-6	0	1	
7	6	0.553	0	0	0	8	0	0	
8	7	0.555	0	0	0	6	0	0	
9	8	0.561	0	0	0	-5	0	0	
10	9	0.564	0	0	0	6	0	0	
11	10	0.565	0	0	0	7	0	0	

(1) R の使い方

まずは、R にデータを読み込みましょう。データの入力方法については、第 9 章で説明しましたので、本資料の 1、2 ページを確認してください。

R でのデータ読み込み

```
> data<- read.csv("10_regression_kyojin.csv", header=TRUE, fileEncoding="CP932")
> head(data)
```

	番号	入場率	休日	伝統の一戦	交流戦	貯金	応援グッズ	野球グッズ	プレゼント
1	1	0.300	0	0	0	5		0	0
2	2	0.309	0	0	0	6		0	1
3	3	0.462	0	0	0	2		0	0
4	4	0.484	0	0	0	5		0	1
5	5	0.488	0	0	0	-6		0	1
6	6	0.553	0	0	0	8		0	0

次に、説明変数の相関係数を求めたいのですが、**相関行列**は cor 関数で作成できます⁴。関数の括弧内にはデータを入力します。データ data の 3 列目から 8 列目が説明変数に相当しますので、「data [3:8]」でデータの範囲を指定しています。

⁴ cor.test 関数を用いると、相関係数が有意であるかを検証できます。

round 関数で出力される数字の桁数を調整できます。「digits」で桁数を指定しますので、小数点第 3 位までの値を出力したい場合、「digits = 3」とします。交流戦と貯金の相関が最も大きく、その値は 0.341 であることを確認できるでしょう。多重共線性の疑いを判断するときに指標として用いる、**分散拡大係数 (VIF)** は後ほど別の関数で出力します。

R での相関行列の作成

```
> # 相関係数
> round(cor(data[3:8]), digits=3)
```

	休日	伝統の一戦	交流戦	貯金	応援グッズプレゼント	野球グッズプレゼント
休日	1.000	0.063	-0.061	0.006	-0.147	0.142
伝統の一戦	0.063	1.000	-0.175	-0.041	-0.087	0.000
交流戦	-0.061	-0.175	1.000	0.341	-0.038	-0.096
貯金	0.006	-0.041	0.341	1.000	0.055	-0.161
応援グッズプレゼント	-0.147	-0.087	-0.038	0.055	1.000	-0.171
野球グッズプレゼント	0.142	0.000	-0.096	-0.161	-0.171	1.000

では、**線形回帰分析**を行ってみましょう。lm 関数で分析を実行できます。今回、モデル 1 にコントロール変数のみ、モデル 2 にすべての変数を投入し、モデル比較を行います。分散分析の aov 関数と同様、「~」の左側には**目的変数**、右側には**説明変数**を入力します。説明変数が 2 つ以上の**重回帰分析**の場合、「+」で説明変数をつなげて書いてください。data という名称で、CSV ファイルを読み込んでいますので、「data = data」となります。

分析結果は summary 関数で出力できます。モデル 1 の情報は「res1」に、モデル 2 の情報は「res2」に入れていますので、「summary (res2)」でモデル 2 の結果を見てみましょう。

一番下に、F 検定の結果が示されています。**p 値**は「2.969e-12」ですので⁵、0.001%水準で有意であることが示されました。**自由度調整済み決定係数**は「Adjusted R-squared」に表示されていますので、モデル 1 の値と見比べて、モデル比較を行ってみましょう。**赤池情報量規準 (AIC)**は以下のように、AIC 関数で出力できます。括弧内にはモデルの情報を入力してください。モデル 1 よりも、モデル 2 の AIC が小さいことが示されました。

回帰係数の推定結果は、summary 関数で出力された「Coefficients」のところを見ればわかります。推定値は「Estimate」、**p 値**は「Pr(> |t|)」のところに記載されています。巨人のデータにおいて、応援グッズプレゼントが入場率に与える効果は 5%水準で有意であることが明らかになりました。これに対して、野球グッズプレゼントが入場率に与える効果は非有意で、野球グッズの配布効果があるとはいえないことも示されました。

⁵ 「e」は指数表示で、 2.969×10^{-12} (= 0.000000000002969)であることを意味します。

R における線形回帰分析の結果

```

> # モデル1:コントロール変数のみ
> # モデル2:すべての変数
> res1<- lm(入場率~休日+伝統の一戦+交流戦+貯金,data=data)
> res2<- lm(入場率~休日+伝統の一戦+交流戦+貯金+応援グッズプレゼント+野球グッズプレゼント,data=data)
> summary(res2)

Call:
lm(formula = 入場率 ~ 休日 + 伝統の一戦 + 交流戦 +
    貯金 + 応援グッズプレゼント + 野球グッズプレゼント,
    data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-0.40082 -0.06842  0.01077  0.06967  0.22922

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.748859   0.016041  46.683 < 2e-16 ***
休日            0.150763   0.019395   7.773 1.68e-12 ***
伝統の一戦      0.067987   0.025128   2.706 0.007691 **
交流戦          0.091220   0.030595   2.981 0.003400 **
貯金           -0.009607   0.002399  -4.005 0.000102 ***
応援グッズプレゼント 0.053983   0.027206   1.984 0.049240 *
野球グッズプレゼント -0.030261   0.021572  -1.403 0.162958
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1117 on 136 degrees of freedom
Multiple R-squared:  0.3794,    Adjusted R-squared:  0.352
F-statistic: 13.86 on 6 and 136 DF,  p-value: 2.969e-12

```

R での AIC の算出

```

> AIC(res1)
[1] -209.2476
> AIC(res2)
[1] -212.3307

```

最後に、**多重共線性の疑い**がないかを確認しましょう。R では、利用者が自由に自作した分析用の関数やツールがパッケージとして公開されています。**VIF** を算出するための **vif** 関数は **car** パッケージに入っています。はじめて、パッケージ内の関数やツールを利用する際、**install.packages** 関数でダウンロードを行います。「” ”」内にパッケージ名 **car** を入力した後、「Japan (Yonezawa)」を指定してください。2 回目以降、同じパッケージを利用する場合には、ダウンロードの作業は必要ありません。

ダウンロードしたパッケージの情報は **library** 関数で呼び出せます。**library (car)** を実行した後、**vif** 関数で **VIF** を出力してみましょう。括弧内にはモデルの情報を入力します。**VIF** を見てみると、その値は 5 よりも小さいため⁶、多重共線性の疑いはないと判断できます。

⁶ Hair, J. F., Black, VV. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate data analysis (7th ed.)*. Upper Saddle River, N. J. : Pearson Prentice Hall.

R での VIF の算出

```
> # 多重共線性
> install.packages("car")
パッケージを 'C:/Users/tadarei/AppData/Local/R/win-library/4.4' 中にインストールします
('lib' が指定されていないため)
--- このセッションで使うために、CRAN のミラーサイトを選んでください ---
URL 'https://ftp.yz.yamagata-u.ac.jp/pub/cran/bin/windows/contrib/4.4/car_3.1-3.zip' を試しています
Content type 'application/zip' length 1541859 bytes (1.5 MB)
downloaded 1.5 MB

パッケージ 'car' は無事に展開され、MD5 サムもチェックされました

ダウンロードされたパッケージは、以下にあります
  C:\Users\tadarei\AppData\Local\Temp\RtmpwF4oQn\downloaded_packages
> library(car)
要求されたパッケージ carData をロード中です
警告メッセージ:
パッケージ 'car' はバージョン 4.4.3 の R の下で造られました
> vif(res1)
      休日 伝統の一戦      交流戦      貯金
1.007350  1.034757  1.169343  1.132713

> vif(res2)
      休日      伝統の一戦      交流戦
1.045384  1.044550  1.181089
      貯金 応援グッズプレゼント 野球グッズプレゼント
1.159965  1.063368  1.075141
```

(2) HAD の使い方

HAD に日本プロ野球の統計データを読み込んでみましょう。データの入力方法については、第 9 章で説明しましたので、本資料の 3、4 ページを確認してください。

HAD のモデリング画面

	A	B	C	D	E	F	G	H	I	J
1	変数名	番号	入場率	休日	伝統の一戦	交流戦	貯金	応援グッズ	野球グッズ	プレゼント
2										
3	分析		選択セルを使用		統制変数を投入		変数を左につめる			● 変数情報
4										○ 回帰分析
5	データ									○ 因子分析
6	シート									
7										
8										
9	使用変数	番号								
10										
300	変数情報	フィルタ	値ラベル	ラベル	@変数	コード				
301	番号									
302	入場率									
303	休日									
304	伝統の一戦									
305	交流戦									
306	貯金									
307	応援グッズ									
308	野球グッズ									
	プレゼント									

はじめに、説明変数の相関係数を求め、**相関行列**としてまとめてみましょう。「使用変数」をクリックして、「データリスト」から休日、伝統の一戦、交流戦、貯金、応援グッズプレゼント、野球グッズプレゼントを選択します。次に、「変数情報」を選択し、「分析」をクリックしてください。「詳細モード」を押すと、以下の画面が現れますので、「相関分析」と「行列表示」にチェックを入れたことを確認し、OK を押しましょう。

相関分析を実行すると、新しいシート「**Corr**」が出現します。相関係数の桁数は Excel「ホーム画面」の「数値」で調整できますので、小数点第 3 位までを表示しました。交流戦と貯金の相関が最も大きく、その値は 0.341 であることを確認できるでしょう。多重共線性の疑いを判断するときの指標である、**分散拡大係数 (VIF)** は後ほど別の形で出力します。

HAD での相関行列の作成

統計分析マクロ HAD

×

データ入力方法

☒ ローデータとして読み込む

☐ クロス表として読み込む

欠損値処理方法

☒ ペアワイズ削除

☐ リストワイズ削除

データの要約

☐ 要約統計量☐ 箱ひげ図

☐ ヒストグラム☐ 度数分布表

☐ カーネル密度☐ 正規性検定

☐ 散布図☐ クロス表

☐ グループ別☐ バブルチャート

☐ 群ごとの統計量☐ 曲線当てはめ

変数間の関連性

☒ 相関分析☐ 共分散行列

☒ 行列表示☐ 主成分分析

☐ 順位相関分析☐ カテゴリカル相関

☐ 行列表示☐ 行列表示

☐ 項目分析☐ 信頼性係数

☐ セルごとの度数☐ 対応分析

差の検定

☐ 平均値の推定

検定値 =

☐ 平均値の差の検定

☒ 対応なし

☐ 対応あり

☐ 順位の差の検定 (ノンパラ検定)

☒ 対応なし

☐ 対応あり

☐ 相関の差の検定

☐ 等分散の検定

マルチレベル分析

☐ 級内相関係数

☐ マルチレベル相関分析

☐ MUML用共分散行列

☐ PW相関分析

☐ 識別可能

☐ ペアワイズ共分散行列

☐ グループごとの回帰直線

☐ 出力を上書きしない

☐ フィルタをオフにする

簡易モード

グラフ設定

OK

キャンセル

	A	B	C	D	E	F	G	H	I
1	rowtype_	varname_	休日	伝統の一戦	交流戦	貯金	応援グッズ	野球グッズ	プレゼント
2	n		143	143	143	143	143	143	
3	corr	休日	1.000	0.063	-0.061	0.006	-0.147	0.142	
4	corr	伝統の一戦	0.063	1.000	-0.175	-0.041	-0.087	0.000	
5	corr	交流戦	-0.061	-0.175	1.000	0.341	-0.038	-0.096	
6	corr	貯金	0.006	-0.041	0.341	1.000	0.055	-0.161	
7	corr	応援グッズ	-0.147	-0.087	-0.038	0.055	1.000	-0.171	
8	corr	野球グッズ	0.142	0.000	-0.096	-0.161	-0.171	1.000	

では、**線形回帰分析**を実行してみます。今、**説明変数**は使用可能な形式になっていますので、「使用変数」から**目的変数**である、入場率を追加してください。「モデリング」画面の「**回帰分析**」を選択し、「モデル保存」横の「**回帰分析**」を選択しましょう。

すべての説明変数を投入したモデル 2 を作成します。モデル比較を行うには、コントロール変数のみを投入したモデル 1 も作る必要がありますが、同じ手続きを実行することになるため、ここでは割愛します。「使用変数」横の「入場率」にカーソルを合わせ、「目的変数

を投入」をクリックします。さらに、休日、伝統の一戦、交流戦、貯金、応援グッズプレゼント、野球グッズプレゼントにカーソルを合わせ、「主効果を全投入」をクリックしてください。「目的変数」と「モデル」に以下の情報が入力されているかを確認しましょう。これで準備は終わりましたので、「分析実行」を押して、結果を見ていきます。

HAD の分析実行画面

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	変数名	番号	入場率	休日	伝統の一戦	交流戦	貯金	応援グッズ	野球グッズプレゼント				
2													
3	分析		選択セルを使用		統制変数を投入		変数を左につめる		<input type="radio"/> 変数情報 <input checked="" type="radio"/> 回帰分析 <input type="radio"/> 因子分析		変数の作成		
4													
5	データシート											データセット	
6													
7													
8													
9	使用変数	番号	入場率	休日	伝統の一戦	交流戦	貯金	応援グッズ	野球グッズプレゼント				
10													
11													
12	重回帰分析		目的変数を投入		主効果を全投入		交互作用を全投入		オプション		分析実行		
13													
14													
15	目的変数→	入場率											
16													
17	モデル→	休日	伝統の一戦	交流戦	貯金	応援グッズ	野球グッズプレゼント						
18													
19													
20													
21	重み→												
22													
23	スライス→				スライス		「階層的投入法		「ステップワイズ		「各ステップを出力しない		「半
24									「媒介分析		「残差得点を出力		「多
25													
26	モデル保存	<input checked="" type="radio"/> 回帰分析	<input type="radio"/> 分散分析	<input type="radio"/> 一般化線形モデル	<input type="radio"/> 一般化線形混合モデル								
27													

線形回帰分析の結果は以下の通りです。分析を実行すると、新しいシート「**Reg**」が出現します。 F 検定の結果、回帰モデルは 0.001%水準で有意であることが示されました。自由度調整済み決定係数は「Adjusted R^2 」に表示されており、0.352 であることがわかります。赤池情報量規準（AIC）は -212.331 と出力されています。モデルを比較するときは、これらの値を用いて、自由度調整済み決定係数の大きさ、AIC の小ささを確かめてください。

HAD での自由度調整済み決定係数および AIC の算出

	A	B	C	D	E	F	G	H	I	J	K
1											
2	重回帰分析				サンプルサイズ = 143						
3											
4	Step1	入場率	<-	休日	伝統の一戦	交流戦	貯金	応援グッズ	野球グッズ	プレゼント	
5											
6											
7	モデル適合										
8											
9			SS	df	MS	F 値	p 値				
10	モデル		1.037	6	0.173	13.857	.000				
11	誤差		1.696	136	0.012						
12	全体		2.733	142							
13											
14											
15		適合指標	R ²	Adjust R ²	F 値	df	p 値	AIC	BIC	CAIC	
16			.379	.352	13.857	6, 136	.000	-212.331	-188.628	---	

次に、**回帰係数**の推定結果を見てみましょう。「係数」と「*p*値」を確認しながら、説明変数が目的変数にどのような影響を及ぼしているかを判断します。分析の結果、応援グッズプレゼントが入場率に与える効果は 5%水準で有意で、野球グッズプレゼントが入場率に与える効果は非有意でした。つまり、応援グッズプレゼントをフリーギフトとして配布することで、試合の入場率が増加している傾向が明らかになりました。

HAD における線形回帰分析の結果

19	回帰係数	目的変数 = 入場率								
20										
21		変数名	係数	標準誤差	95%下限	95%上限	<i>t</i> 値	df	<i>p</i> 値	
22		切片	0.749	0.016	0.717	0.781	46.683	136	.000	
23		休日	0.151	0.019	0.112	0.189	7.773	136	.000	**
24		伝統の一戦	0.068	0.025	0.018	0.118	2.706	136	.008	**
25		交流戦	0.091	0.031	0.031	0.152	2.981	136	.003	**
26		貯金	-0.010	0.002	-0.014	-0.005	-4.005	136	.000	**
27		応援グッズプレゼント	0.054	0.027	0.000	0.108	1.984	136	.049	*
28		野球グッズプレゼント	-0.030	0.022	-0.073	0.012	-1.403	136	.163	

最後に、**多重共線性の疑い**がないかを確認しましょう。HAD では、「Reg」のシートにおける一番下の表に VIF の値が出力されます。VIF の最大値は 1.181 で、いずれも 5 より小さいため⁷、多重共線性の疑いはないと判断できます。

HAD での VIF の算出

32	標準化係数	目的変数 = 入場率			
33					
34	変数名	入場率	95%下限	95%上限	VIF
35	休日	.537 **	0.400	0.673	1.045
36	伝統の一戦	.187 **	0.050	0.323	1.045
37	交流戦	.219 **	0.074	0.364	1.181
38	貯金	-.291 **	-0.435	-0.148	1.160
39	応援グッズプレゼント	.138 *	0.000	0.276	1.063
40	野球グッズプレゼント	-.098	-0.237	0.040	1.075
41	R^2	.379 **			
42					
43		** $p < .01$, * $p < .05$, + $p < .10$			

⁷ Hair, J. F., Black, V. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate data analysis (7th ed.)*. Upper Saddle River, N. J. : Pearson Prentice Hall.

第 11 章 因子分析

第 11 章では、質問紙調査で収集した統計データ (11_factor.csv) を利用して、探索的因子分析と線形回帰分析を実行します。CSV ファイルの一部を以下に掲載しておきますので、確認してください。

分析データ (一部抜粋)

	A	B	C	D	E	F	G	H	I	J
1	番号	価値観1	価値観2	価値観3	価値観4	価値観5	価値観6	価値観7	価値観8	価値観9
2	1	5	5	5	5	5	5	4	4	4
3	2	4	4	4	3	3	3	4	4	4
4	3	4	4	5	5	5	5	1	1	1
5	4	4	4	4	5	4	5	5	5	2
6	5	4	4	4	4	2	5	4	2	1
7	6	4	5	4	2	3	2	4	4	5
8	7	4	4	5	4	3	1	3	4	5
9	8	5	5	5	4	3	3	4	4	4
10	9	5	5	3	3	3	2	4	4	3
11	10	4	4	4	4	2	2	2	3	1

注：価値観 9 の右側にもデータが入っています。

(1) R の使い方

まずは、R にデータを読み込みましょう。データの入力方法については、第 9 章で説明しましたので、本資料の 1、2 ページを確認してください。

R でのデータ読み込み

```
> data<- read.csv("11_factor.csv", header=TRUE, fileEncoding="CP932")
> head(data)
```

	番号	価値観1	価値観2	価値観3	価値観4	価値観5	価値観6	価値観7	価値観8	価値観9
1	1	5	5	5	5	5	5	4	4	4
2	2	4	4	4	3	3	3	4	4	4
3	3	4	4	5	5	5	5	1	1	1
4	4	4	4	4	5	4	5	5	5	2
5	5	4	4	4	4	2	5	4	2	1
6	6	4	5	4	2	3	2	4	4	5

	年齢	性別	月収	クレジットカード	QRコード決済
1	21	1	5		5
2	20	1	4		1
3	20	1	4		5
4	21	1	4		3
5	20	0	5		4
6	20	1	2		5

データは左から順に、識別用の番号、決済手段に関する価値観（価値観 1～9）、回答者の年齢、性別、月収、クレジットカードと QR コード決済の利用頻度です。分析の目的は、価値観の 9 変数から潜在的な因子を抽出し、変数を要約することです。因子分析に適したデータであるかを確認するため⁸、価値観の変数の**相関行列**を作成します。

第 10 章で説明したように、**相関行列**は `cor` 関数で作成できます。括弧内にデータを入力しますが、データ `data` の 2 列目から 10 列目が価値観の変数に相当しますので、「`data[2:10]`」でデータの範囲を指定しています。`round` 関数で出力される数字の桁数を調整できますが、小数点第 3 位までの値を出力したいため、「`digits = 3`」としました。価値観 1～3、価値観 4～6、価値観 7～9 における相関係数が大きいことを確認できます。

R での相関行列の作成

```
> # 相関係数
> round(cor(data[2:10]), digits=3)
```

	価値観1	価値観2	価値観3	価値観4	価値観5	価値観6	価値観7	価値観8	価値観9
価値観1	1.000	0.698	0.453	0.045	0.059	0.191	0.212	0.196	0.128
価値観2	0.698	1.000	0.496	0.100	0.114	0.185	0.267	0.267	0.201
価値観3	0.453	0.496	1.000	0.175	0.196	0.166	0.124	0.167	0.137
価値観4	0.045	0.100	0.175	1.000	0.729	0.515	0.155	0.216	0.234
価値観5	0.059	0.114	0.196	0.729	1.000	0.582	0.202	0.195	0.285
価値観6	0.191	0.185	0.166	0.515	0.582	1.000	0.329	0.219	0.197
価値観7	0.212	0.267	0.124	0.155	0.202	0.329	1.000	0.651	0.475
価値観8	0.196	0.267	0.167	0.216	0.195	0.219	0.651	1.000	0.587
価値観9	0.128	0.201	0.137	0.234	0.285	0.197	0.475	0.587	1.000

因子分析に適したデータであることを確認できたならば、因子数を決めましょう⁹。ここでは、**カイザー・ガットマン基準**に依拠し、1 以上の**固有値**の数を確認します。`eigen` 関数で固有値と固有ベクトルを算出できますが、固有値のみを取り出すため、「`$values`」の指示を与えました。関数の括弧内には相関行列を入力するため、「`cor (data [2:10])`」と記入しています。出力結果から、1 以上の固有値の数は 3 であることがわかりました。

R での固有値の算出

```
> # 因子数の決定
> eigen(cor(data[2:10]))$values
[1] 3.2716007 1.7804169 1.4613442 0.6838529 0.5253772 0.4253073 0.3177214
[8] 0.2888151 0.2455645
```

では、**因子分析**を行ってみましょう。カイザー・ガットマン基準より、因子数は「3」としました。因子分析には `psych` パッケージの `fa` 関数を用います。本資料 10 ページでも説明

⁸ `psych` パッケージの `KMO` 関数を用いると、サンプリング妥当性尺度の `MSA` (measure of sampling adequacy)、`KMO` (Kaiser-Meyer-Olkin) を算出できます。

⁹ `psych` パッケージの `fa.parallel` 関数を用いると、平行分析を実行できます。

したように、パッケージの関数やツールをはじめて使うときは、`install.packages` 関数でインストールを行ってください。`library` 関数でパッケージの情報を呼び出しています。

`fa` 関数の括弧内にはデータ、因子数、推定方法、軸の回転方法を記載します。「`nfactors = 3`」は因子数を 3 とすること、「`fm = "ml"`」は**最尤法**で推定すること¹⁰、「`rotate = "Promax"`」は**プロマックス回転**を行うこと¹¹を意味します。推定結果は `print` 関数で出力し、**因子負荷量**を小数点第 3 位まで表示するため、「`digits = 3`」に設定しました。

推定結果の見方ですが、「ML1~3」は各変数の因子負荷量、「h2」は共通性、「u2」は独自性、「com」は複雑性を表します。その下の「Proportion Var」は**寄与率**、「Cumulative Var」は**累積寄与率**を示しています。因子負荷量を見ると、左から順に、価値観 4~6 は因子 1 (ML1: 節約因子)、価値観 7~9 は因子 2 (ML3: 支払管理因子)、価値観 1~3 は因子 3 (ML2: 利便性因子) を表していると解釈できるでしょう。3 因子による累積寄与率は 0.604 であり、共通性は 1 を超えていませんので、うまく推定できていることも確認できます。

R における因子分析の結果

```
> # 因子分析の実行
> library(psych)
> res<- fa(data[2:10], nfactors=3, fm="ml", rotate="Promax")
> print(res, digits=3)
Factor Analysis using method = ml
Call: fa(r = data[2:10], nfactors = 3, rotate = "Promax", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
```

	ML1	ML3	ML2	h2	u2	com
価値観1	-0.079	-0.025	0.842	0.668	0.332	1.02
価値観2	-0.048	0.040	0.856	0.738	0.262	1.01
価値観3	0.121	-0.045	0.558	0.341	0.659	1.11
価値観4	0.810	-0.016	-0.061	0.624	0.376	1.01
価値観5	0.962	-0.066	-0.058	0.857	0.143	1.02
価値観6	0.603	0.047	0.086	0.426	0.574	1.05
価値観7	-0.021	0.721	0.053	0.537	0.463	1.01
価値観8	-0.082	0.928	-0.007	0.803	0.197	1.02
価値観9	0.096	0.632	-0.027	0.444	0.556	1.05

	ML1	ML3	ML2
SS loadings	1.926	1.761	1.750
Proportion Var	0.214	0.196	0.194
Cumulative Var	0.214	0.410	0.604
Proportion Explained	0.354	0.324	0.322
Cumulative Proportion	0.354	0.678	1.000

最後に、3 つの因子がクレジットカードと QR コード決済の利用頻度に及ぼす影響を**線形回帰分析**で検討してみましょう。分析モデルの情報を格納した「res」の「scores」に**因子得点**のデータが入っていますので、「res\$scores」で因子得点を出力可能です。`cbind` 関数を用い

¹⁰ 最小二乗法の場合は「minres」、主因子法の場合は「pa」と表記します。

¹¹ 回転なしの場合は「none」、バリマックス回転の場合は「varimax」と表記します。

と、データを列方向に拡張できますので、「`cbind(data,res$scores)`」により、データ `data` と因子得点のデータを結合しました。新しいデータ `ndat` の概要は以下の通りです。

R でのデータ整理

```
> ndat<- cbind(data,res$scores)
> head(ndat)
```

番号	価値観1	価値観2	価値観3	価値観4	価値観5	価値観6	価値観7	価値観8	価値観9	年齢
1	1	5	5	5	5	5	4	4	4	21
2	2	4	4	4	3	3	3	4	4	20
3	3	4	4	5	5	5	5	1	1	20
4	4	4	4	4	5	4	5	5	5	21
5	5	4	4	4	4	2	5	4	2	20
6	6	4	5	4	2	3	2	4	4	20

	性別	月収	クレジットカード	QRコード決済	ML1	ML3	ML2	
1	1	5		5	5	1.5393815	0.4829473	0.9352840
2	1	4		1	1	-0.1115643	0.2694070	-0.4460710
3	1	4		2	5	1.2678694	-2.0697022	-0.3326786
4	1	4		3	4	0.9284387	0.8619193	-0.3188745
5	0	5		5	4	-0.4247642	-1.1199955	-0.4465252
6	1	2		2	5	-0.3113072	0.3940777	0.1138253

この結合したデータ `ndat` と `lm` 関数を用いて、線形回帰分析を実行します。回帰分析のやり方については、第 10 章を確認してください。クレジットカードの利用頻度を**目的変数**、年齢、性別、月収、因子得点を**説明変数**としたモデルの推定結果を以下に示します。

R における線形回帰分析の結果

```
> res1<- lm(クレジットカード~年齢+性別+月収+ML1+ML3+ML2,data=ndat)
> summary(res1)
```

Call:
lm(formula = クレジットカード ~ 年齢 + 性別 + 月収 + ML1 + ML3 + ML2, data = ndat)

Residuals:

	Min	1Q	Median	3Q	Max
	-3.6019	-1.3975	0.1007	1.3664	3.6655

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-5.23748	1.26776	-4.131	4.58e-05	***
年齢	0.38465	0.06600	5.828	1.34e-08	***
性別	0.17202	0.17641	0.975	0.33022	
月収	0.17394	0.07078	2.457	0.01451	*
ML1	-0.12907	0.09977	-1.294	0.19668	
ML3	0.27468	0.10323	2.661	0.00818	**
ML2	0.06083	0.09825	0.619	0.53628	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.506 on 329 degrees of freedom
Multiple R-squared: 0.1516, Adjusted R-squared: 0.1361
F-statistic: 9.796 on 6 and 329 DF, p-value: 6.119e-10

(2) HAD の使い方

HAD に質問紙調査で収集した統計データを読み込んでみましょう。データの入力方法については、第 9 章で説明しましたので、本資料の 3、4 ページを確認してください。

HAD のモデリング画面

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	変数名	番号	価値観1	価値観2	価値観3	価値観4	価値観5	価値観6	価値観7	価値観8	価値観9	年齢	性別
2													
3	分析		選択セルを使用		統制変数を投入		変数を左につめる		<input checked="" type="radio"/> 変数情報		変数の作成		
4									<input type="radio"/> 回帰分析				
5									<input type="radio"/> 因子分析		データセット		
6	データ												
7	シート												
8													
9	使用変数	番号											
10													
300	変数情報	フィルタ	値ラベル	ラベル	@変数	コード							
301	番号												
302	価値観1												
303	価値観2												
304	価値観3												
305	価値観4												
306	価値観5												
307	価値観6												
308	価値観7												
309	価値観8												
310	価値観9												
311	年齢												
312	性別												
313	月収												
314	クレジットカード												
315	QRコード決済												

データは左から順に、識別用の番号、決済手段に関する価値観（価値観 1～9）、回答者の年齢、性別、月収、クレジットカードと QR コード決済の利用頻度です。分析の目的は、価値観の 9 変数から潜在的な因子を抽出し、変数を要約することです。

因子分析に適したデータであるかを確認するため、価値観の変数の**相関行列**を作成してみましょう。第 10 章で説明したように、「使用変数」をクリックして、「データリスト」から、価値観 1～9 を選択します。次に、「変数情報」を選択し、「分析」をクリックしてください。「詳細モード」で「相関分析」と「行列表示」にチェックを入れて、OK を押しましょう。相関分析を実行すると、新しいシート「**Corr**」が出現します。相関係数の桁数は Excel「ホーム画面」の「数値」で調整し、小数点第 3 位までを表示しました。相関行列を見ると、価値観 1～3、価値観 4～6、価値観 7～9 における相関係数が高いことを確認できます。

HAD での相関行列の作成

	A	B	C	D	E	F	G	H	I	J	K
1	rowtype_	varname_	価値観1	価値観2	価値観3	価値観4	価値観5	価値観6	価値観7	価値観8	価値観9
2	n		336	336	336	336	336	336	336	336	336
3	corr	価値観1	1.000	0.698	0.453	0.045	0.059	0.191	0.212	0.196	0.128
4	corr	価値観2	0.698	1.000	0.496	0.100	0.114	0.185	0.267	0.267	0.201
5	corr	価値観3	0.453	0.496	1.000	0.175	0.196	0.166	0.124	0.167	0.137
6	corr	価値観4	0.045	0.100	0.175	1.000	0.729	0.515	0.155	0.216	0.234
7	corr	価値観5	0.059	0.114	0.196	0.729	1.000	0.582	0.202	0.195	0.285
8	corr	価値観6	0.191	0.185	0.166	0.515	0.582	1.000	0.329	0.219	0.197
9	corr	価値観7	0.212	0.267	0.124	0.155	0.202	0.329	1.000	0.651	0.475
10	corr	価値観8	0.196	0.267	0.167	0.216	0.195	0.219	0.651	1.000	0.587
11	corr	価値観9	0.128	0.201	0.137	0.234	0.285	0.197	0.475	0.587	1.000

では、**因子分析**を実行してみましょう。「モデリング」画面の「**因子分析**」を選択します。分析にあたり、因子数を決める必要がありますが、ここでは**カイザー・ガットマン基準**に依拠します。HAD では、「因子数」の「固有値 1 以上まで」にチェックを入れると、自動的に分析が実行されます。「抽出法」は「**最尤法**」、「回転法」は「**斜交回転（プロマックス回転）**」としました。後の回帰分析で利用するため、「**因子得点**」にもチェックを入れてください。これで準備は終わりましたので、「分析実行」を押して、結果を見ていきます。

HAD の分析実行画面

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	変数名	番号	価値観1	価値観2	価値観3	価値観4	価値観5	価値観6	価値観7	価値観8	価値観9	年齢	性別
2													
3	分析		選択セルを使用		統制変数を投入		変数を左につめる			<input type="radio"/> 変数情報		変数の作成	
4										<input type="radio"/> 回帰分析			
5	データシート									<input checked="" type="radio"/> 因子分析		データセット	
6													
7													
8													
9	使用変数	番号	価値観1	価値観2	価値観3	価値観4	価値観5	価値観6	価値観7	価値観8	価値観9		
10													
28	因子分析		最尤プロマックス		主成分法					オプション		分析実行	
29													
30													
31													
32	因子数→	1	<input checked="" type="checkbox"/> 固有値1以上まで		スクリープロット								
33													
34	抽出法→	<input checked="" type="radio"/> 最尤法	<input type="radio"/> 最小二乗法	<input type="radio"/> 反復主因子法	<input type="radio"/> 主成分法	<input type="radio"/> カタリカル							
35													
36	回転法→	<input checked="" type="radio"/> 斜交回転	<input type="radio"/> 直交回転	<input type="radio"/> 回転なし	<input type="checkbox"/> プロクラス								
37													
38	得点→	<input checked="" type="checkbox"/> 因子得点	<input type="checkbox"/> 尺度得点										
39													
40	出力→	<input checked="" type="checkbox"/> サイズでソート	<input type="checkbox"/> 相関行列		<input type="checkbox"/> 項目反応理論								
41													
42													
43	モデル保存	<input checked="" type="radio"/> 因子分析	<input type="radio"/> クラス分析	<input type="radio"/> 数量化分析	<input type="radio"/> 構造方程	<input type="radio"/> ファイル							
44													

因子分析の結果は以下の通りです。分析を実行すると、新しいシート「**Factor**」と「**ScoreF**」が出現します。前者には**因子負荷量**の推定結果、後者には**因子得点**が掲載されています。因子負荷量を見れば、価値観 1～9 は潜在的な 3 つの因子で要約できることがわかります。左から順に、価値観 7～9 は因子 1（Factor 1: 支払管理因子）、価値観 4～6 は因子 2（Factor 2: 節約因子）、価値観 1～3 は因子 3（Factor 3: 利便性因子）を表していると解釈できるでしょう。共通性の値は 1 を超えていませんので、うまく推定できていることも確認できました。

HAD における因子分析の結果

15	項目	Factor1	Factor2	Factor3	共通性
16	価値観8	.921	-.055	-.025	.803
17	価値観7	.717	.000	.041	.537
18	価値観9	.632	.110	-.032	.444
19	価値観5	-.032	.939	-.019	.857
20	価値観4	.013	.791	-.029	.624
21	価値観6	.069	.595	.108	.426
22	価値観2	.043	-.021	.847	.738
23	価値観1	-.023	-.053	.832	.668
24	価値観3	-.038	.134	.559	.341
25					
26	因子寄与	2.206	2.186	2.016	
27					
28	適合度	乖離度 =	0.093	CFI =	.983
29		χ^2 値 =	30.694	RMSEA =	.069
30		DF =	12	AIC =	79.238
31		p =	.002	BIC =	170.849

最後に、3 つの因子がクレジットカードと QR コード決済の利用頻度に及ぼす影響を**線形回帰分析**で検討してみましょう。シート「**ScoreF**」のデータを「データ」画面に貼り付けて、再度データの読み込みを行ってください。「使用変数」をクリックし、「データリスト」から、年齢、性別、月収、クレジットカード、QR コード決済、Factor 1～3 を追加します。

これらの追加したデータを用いて、線形回帰分析を実行します。回帰分析のやり方については、第 10 章を確認してください。クレジットカードの利用頻度を**目的変数**、年齢、性別、月収、因子得点を**説明変数**としたモデルの推定結果を以下に示します。

HAD の分析実行画面

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	変数名	番号	価値観1	価値観2	価値観3	価値観4	価値観5	価値観6	価値観7	価値観8	価値観9	年齢	性別
2													
3	分析		選択セルを使用		統制変数を投入		変数を左につめる			<input type="radio"/> 変数情報	変数の作成		
4									<input checked="" type="radio"/> 回帰分析				
5										<input type="radio"/> 因子分析	データセット		
6	データシート												
7													
8	使用変数	番号	年齢	性別	月収	クレジットカード	Factor1	Factor2	Factor3				
9													
10													
11	重回帰分析		目的変数を投入		主効果を全投入		交互作用を全投入			オプション		分析実行	
12													
13													
14	目的変数→	クレジットカード											
15													
16	モデル→	年齢	性別	月収	Factor1	Factor2	Factor3						
17													
18													
19													
20													
21	重み→												
22													
23	スライス→				スライス		<input type="checkbox"/> 階層的投入法		<input type="checkbox"/> ステップワイズ		<input type="checkbox"/> 各ステップを出力しない		<input type="checkbox"/> 半
24									<input type="checkbox"/> 媒介分析		<input type="checkbox"/> 残差得点を出力		<input type="checkbox"/> 多
25													
26	モデル保存	<input checked="" type="radio"/> 回帰分析 <input type="radio"/> 分散分析 <input type="radio"/> 一般化線形モデル <input type="radio"/> 一般化線形混合モデル											
27													

HAD における線形回帰分析の結果

7	モデル適合												
8													
9			SS	df	MS	F 値	p 値						
10	モデル	133.223	6	22.204	9.796	.000							
11	誤差	745.703	329	2.267									
12	全体	878.926	335										
13													
14													
15	適合指標	R ²	Adjust R ²	F 値	df	p 値	AIC	BIC	CAIC				
16		.152	.136	9.796	6, 329	.000	1237.391	1267.928	---				
17													
18													
19	回帰係数	目的変数 = クレジットカード											
20													
21	変数名	係数	標準誤差	95%下限	95%上限	t 値	df	p 値					
22	切片	-5.237	1.268	-7.731	-2.744	-4.131	329	.000					
23	年齢	0.385	0.066	0.255	0.514	5.828	329	.000 **					
24	性別	0.172	0.176	-0.175	0.519	0.975	329	.330					
25	月収	0.174	0.071	0.035	0.313	2.457	329	.015 *					
26	Factor1	0.269	0.101	0.070	0.469	2.655	329	.008 **					
27	Factor2	-0.117	0.096	-0.307	0.072	-1.216	329	.225					
28	Factor3	0.051	0.098	-0.141	0.243	0.523	329	.601					

第 12 章 クラスター分析

第 12 章では、「家計調査」の統計データ (12_cluster.csv) を利用して、クラスター分析を実行します。CSV ファイルの一部を以下に掲載しておきますので、確認してください。

分析データ (一部抜粋)

	A	B	C	D	E	F	G
1	都市名	米	パン	生うどん	パスタ	中華麺	もち
2	静岡市	27075	32155	3084	1449	5022	1601
3	新潟市	26224	34151	2800	1665	5185	2159
4	那覇市	25527	28337	1953	1066	4927	1193
5	長崎市	25454	33877	2794	1264	4330	1457
6	浜松市	25377	33136	3344	1316	4621	1483
7	福井市	25272	30765	3796	1136	4159	2786
8	堺市	24983	37575	4164	1179	4170	1776
9	富山市	24715	33217	3642	1300	5163	2786
10	札幌市	24470	28196	3199	1354	4216	1613
11	山形市	23953	26166	4580	1321	6030	1836

(1) R の使い方

まずは、R にデータを読み込みましょう。データの入力方法については、第 9 章で説明しましたので、本資料の 1、2 ページを確認してください。

R でのデータ読み込み

```
> data<- read.csv("12_cluster.csv", header=TRUE, fileEncoding="CP932")
> head(data)
  都市名   米   パン 生うどん_そば パスタ 中華麺   もち
1 静岡市 27075 32155      3084   1449   5022 1601
2 新潟市 26224 34151      2800   1665   5185 2159
3 那覇市 25527 28337      1953   1066   4927 1193
4 長崎市 25454 33877      2794   1264   4330 1457
5 浜松市 25377 33136      3344   1316   4621 1483
6 福井市 25272 30765      3796   1136   4159 2786
```

続いて、データを整理していきます。データの 1 列目は都市名であるため、分析に使用するのは 2 列目～7 列目です。「data [-1]」の「-1」は 1 列目を取り除くことを意味します。それらのデータを ndat と命名しました。

rownames 関数は行名、colnames 関数は列名を取得する関数で、括弧内にデータを入力します。data の 1 列目の都市名を行名に設定したいため、「rownames (ndat)」に「data [,1]」を代入する形で、行名を変更しました。「colnames (ndat)」では、列名を確認しています。

R でのデータ整理

```
> # データの整理
> ndat<- data[-1]
> rownames(ndat)<- data[,1]
> colnames(ndat)
[1] "米"          "パン"          "生うどん_そば" "パスタ"          "中華麺"          "もち"
```

以上の手続きが完了したら、クラスター分析を実行するためのデータを作成します。ここでは、データを標準化した後、ユークリッド距離を算出してみます。scale 関数を用いると、データを標準化できます。括弧内にはデータを入力してください。分析時に混同しないよう、標準化したデータは新しく ndat.s と命名しました。

ユークリッド距離は dist 関数で求められます。標準化するときと同様、括弧内にはデータを入力します。分析時のことを考慮し、距離のデータは ndat.d と命名しました。ユークリッド距離が dist 関数のデフォルトで設定されていますが、引数「method」の設定を変えることで、マンハッタン距離やチェビシェフ距離も算出可能です¹²。

R でのデータ標準化と距離の算出¹³

```
> # データ標準化とユークリッド距離の算出
> ndat.s<- scale(ndat)
> ndat.d<- dist(ndat.s)
> ndat.d
```

	静岡市	新潟市	那覇市	長崎市	浜松市	福井市	堺市	富山市	札幌市
新潟市	2.0721841								
那覇市	3.2921147	4.8557544							
長崎市	1.7438244	3.2110651	2.5978499						
浜松市	1.2830410	2.9462715	3.0462274	0.9880812					
福井市	3.9116645	4.2266074	4.9937765	3.8218352	3.6009844				
堺市	3.1468230	4.0927765	4.6829631	2.5159045	2.1936837	3.2712363			
富山市	3.3224468	3.0412340	5.1215918	3.7257952	3.3540074	1.8478264	3.2954080		
札幌市	1.9529284	3.3042663	2.9376531	1.9432596	1.6540468	3.4159196	3.3247834	3.5822382	
山形市	3.4909584	4.4210580	4.8162203	4.3332608	3.4988588	4.0681043	4.3415818	3.6456190	3.3015265
金沢市	4.1561007	3.3574234	6.3045196	4.6299529	4.2162160	2.6498926	3.8166673	1.3033781	4.4953706
和歌山市	3.1703700	4.3014324	3.6053248	1.8898445	2.0421438	3.2135981	1.5233590	3.4787439	2.8943454
福島市	2.0153333	3.3553715	3.1906912	2.4504133	1.9451438	4.4428735	3.8563561	4.1748479	1.3522676
大阪市	3.0090092	4.1328670	4.0867261	2.0580048	1.8109272	3.5569676	1.1276655	3.6046284	2.8006165
相模原市	2.4897099	2.3147016	4.8529596	2.7480471	2.2775834	3.2068124	2.6352687	2.4875969	2.5993552
盛岡市	3.0504459	4.1348142	3.5735399	3.5450558	3.0230548	5.1018903	4.4435949	4.3188456	3.3442445
京都市	3.6002424	3.6856005	5.7146036	3.3913028	2.9662786	3.7202997	1.8305838	3.1354667	3.8927050
前橋市	3.2818426	4.2975469	4.8736948	3.1782280	2.4200916	3.9014550	2.3145084	3.8011202	2.8255097
鹿児島市	3.1428665	4.2566703	2.9727293	2.1977454	2.4420980	4.1751635	3.7125328	4.5166138	1.6758151
北九州市	3.3632045	4.2800602	3.6125081	2.3200295	2.4898231	3.3546675	3.0478336	3.9248236	1.9755839

では、階層的クラスター分析を行ってみましょう。hclust 関数で階層的クラスター分析を実行できます。括弧内において、集団の形成方法を指定しますが、「method = "ward.D"」はウォード法で集団を形成していくという指示を与えていることになります。

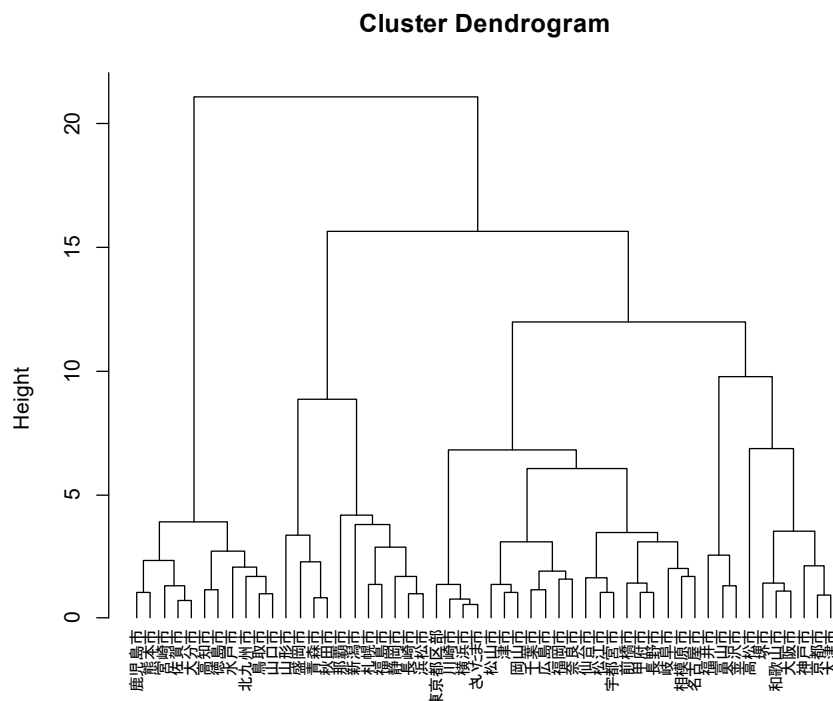
¹² マンハッタン距離を求めるときは「dist (ndat.s,method = "manhattan")」、チェビシェフ距離を求めるときは「dist (ndat.s,method = "maximum")」と入力します。

¹³ 「ndat.d」で出力したユークリッド距離の値は一部です。

集団の導出プロセスを示した、**デンドログラム（樹形図）**は plot 関数で作成できます。括弧内には、分析モデルの情報を表す「res」および図の詳細な設定を入力します。引数「cex」は文字サイズに関する指示で、デフォルトは「1」に設定されています。また、引数「hang」は図の描画に関する指示で、全体を揃えて表示するため、「-1」としました。

R でのデンドログラムの作成

```
> # 階層的クラスター分析  
> res<- hclust(ndat.d, method="ward.D")  
> plot(res, cex=0.7, hang=-1)
```



最後に、クラスターを作成し、それらの特徴量を算出します。上のデンドログラムを確認した結果、今回は3つの集団で52都市を分類してみようと思います。cutree関数を用いて、集団の分割を行いましょ。分析モデルの情報を表す「res」を入力後、括弧内の「k」でクラスターの数を設定します。3つに分割するため、「k=3」と指定してください。

分析結果は「cluster」に格納しました。print関数は出力の指示、for関数は繰り返しの指示を与えるものです。ここで、iはクラスター番号を表します。3つのクラスターを順番に処理しますが、「cluster [cluster == i]」はクラスター番号が i の要素を抜き出し、その都市名を出力するという指示を与えていることになります。

都市名とクラスター番号を出力できたならば、それらの特徴量を求めてみましょう。小計計算は aggregate 関数で実行できます。標準化したデータ ndat.s を用いて、クラスターごと

の平均値を算出するため、「aggregate (ndat.s, list (group = cluster) , mean)」と入力します。引数「list」は集団に関する指示で、分析結果が格納されている「cluster」を指定しました。

出力された値を見ると、クラスター1 は米と中華麺の消費額が多い地域、クラスター2 はパン、生うどん・そば、パスタ、もちの消費額が多い地域、クラスター3 は主食の消費額が少ない地域であることが示されました。

R でのクラスター作成結果

```
> # 3つのクラスターを作成
> cluster<- cutree(res, k=3)
> for(i in 1:3){
+ print(cluster[cluster==i])
+ }
静岡市 新潟市 那覇市 長崎市 浜松市 札幌市 山形市 福島市 盛岡市 青森市 秋田市
      1      1      1      1      1      1      1      1      1      1      1
      福井市      堺市      富山市      金沢市      和歌山市      大阪市      相模原市      京都市
      2      2      2      2      2      2      2      2      2      2
      前橋市      大津市      甲府市      岐阜市      横浜市      長野市      川崎市      名古屋市
      2      2      2      2      2      2      2      2      2      2
さいたま市      千葉市      福岡市      奈良市      仙台市      松江市 東京都区部      神戸市
      2      2      2      2      2      2      2      2      2      2
      広島市      宇都宮市      松山市      津市      高松市      岡山市
      2      2      2      2      2      2      2      2
鹿児島市 北九州市 熊本市 宮崎市 佐賀市 鳥取市 高知市 大分市 徳島市 山口市
      3      3      3      3      3      3      3      3      3      3
      水戸市
      3
> aggregate(ndat.s, list(group=cluster), mean)
  group      米      パン 生うどん_そば      パスタ      中華麺      もち
1     1  1.0236283 -0.7620511  -0.4665405 -0.04788861  1.0327384 -0.4683029
2     2 -0.1760893  0.5521083   0.4086236  0.29723485  0.1099628  0.4272941
3     3 -0.5433848 -0.7436988  -0.6478875 -0.76275190 -1.3326369 -0.6970447
```

(2) HAD の使い方

HAD に「家計調査」の統計データを読み込んでみましょう。データの入力方法については、第9章で説明しましたので、本資料の3、4ページを確認してください。

まず、「使用変数」をクリックして、「データリスト」から米、パン、生うどん_そば、パスタ、中華麺、もちを選択します。次に、データを標準化するため、「変数の作成」ボタンから「尺度変換」メニューの「標準化得点」にチェックを入れ、OK を押します。新しいシート「Scale」が出現しますので、そのデータをコピーして、「データ」画面に貼り付けてください。標準化後のデータの変数名には「_s」が追加されています。再度、データの読み込みを行います。「データリスト」から米_s、パン_s、生うどん_そば_s、パスタ_s、中華麺_s、もち_sを選択し、「使用変数」として指定しましょう。これで分析の準備は完了です。

HAD のモデリング画面

	A	B	C	D	E	F	G	H	I	J
1	変数名	都市名	米	パン	生うどん_そば	パスタ	中華麺	もち		
2										
3	分析		選択セルを使用		統制変数を投入		変数を左につめる		<input checked="" type="radio"/> 変数情報	
4									<input type="radio"/> 回帰分析	
5	データ								<input type="radio"/> 因子分析	
6	シート									
7										
8										
9	使用変数	都市名								
10										
300	変数情報	フィルタ	値ラベル	ラベル	@変数	コード				
301	都市名									
302	米									
303	パン									
304	生うどん_そば									
305	パスタ									
306	中華麺									
307	もち									

HAD での標準化得点の算出

変数の作成： HAD ×

変数の合成

尺度変換

数値変換

値の再割り当て

ダミー変数

・各変数の平均と標準偏差を変換します

☒ 標準化得点 (変数名_s)

平均値= 標準偏差=

☐ 中心化得点 (変数名_c) 平均値を0にします

☐ 順位化得点 (変数名_rs)

・各変数を集団平均で中心化します (マルチレベルデータ)

☐ 集団平均で中心化 (変数名_g) 集団平均で中心化します

☐ 集団平均値 (変数名_m) 集団平均を算出します

☐ フィルタをオフにする

☐ 出力を上書きしない

OK

キャンセル

では、**階層的クラスター分析**を実行してみましょう。モデリング画面の「因子分析」を選択した後、「モデル保存」横の「**クラスター分析**」をクリックします。分析の目的は、主食の消費額から 52 都市を分類することです。ほとんどの場合、最適なクラスター数は事前にわかりませんので、「クラスタ数」は「指定せずに分析」します。「分類方法」の「**階層的**」と「出力」の「**デンドログラム**」にチェックを入れ、「分析実行」を押してください。

なお、集団の形成方法については、**ウォード法**がデフォルトで設定されています。「オプション」画面において、ウォード法以外の集団形成方法を選択することも可能です。

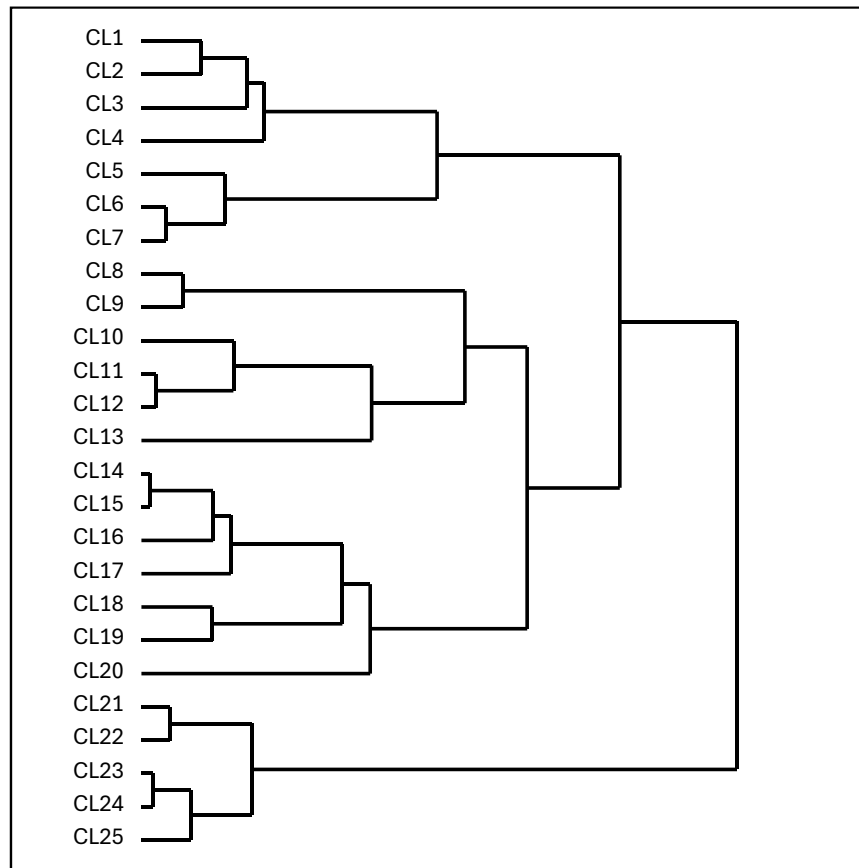
HAD の分析実行画面

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	変数名	都市名	米	パン	生うどん_そば	バスタ	中華麺	もち	米_s	パン_s	生うどん_そば	バスタ_s	中華麺
2													
3	分析		選択セルを使用		統制変数を投入		変数を左につめる		<input type="radio"/> 変数情報		変数の作成		
4								<input type="radio"/> 回帰分析					
5	データシート								<input checked="" type="radio"/> 因子分析		データセット		
6													
7													
8	使用変数	都市名	米_s	パン_s	生うどん_そば	バスタ_s	中華麺_s	もち_s					
9													
10													
28													
29	クラスタ分析		回答者を分類		変数を分類					オプション		分析実行	
30	ウォード法												
31													
32	クラスタ数→	2	<input checked="" type="checkbox"/> 指定せずに分析										
33													
34	分類方法→	<input checked="" type="radio"/> 階層的	<input type="radio"/> 非階層的	<input type="radio"/> 混合分布	<input type="radio"/> 潜在クラス	<input type="radio"/> 潜在ランク							
35													
36	分析対象→	<input checked="" type="radio"/> 回答者	<input type="radio"/> 変数	<input type="radio"/> 距離行列									
37													
38	得点→	<input type="checkbox"/> 所属クラス	<input type="checkbox"/> クラスタ平均値										
39													
40	出力→	<input checked="" type="checkbox"/> デンドログラム	<input type="checkbox"/> クラスタの特徴	<input type="checkbox"/> 多次元尺度法									
41													
42													
43	モデル保存	<input type="radio"/> 因子分析	<input checked="" type="radio"/> クラスタ分析	<input type="radio"/> 数量化分析	<input type="radio"/> 構造方程式モデル								
44													

次に、作成された**デンドログラム**（**樹形図**）を見てみましょう。この図を確認し、クラスター数を決定することになりますが、今回は3つの集団に分割することにしました。

再度、「モデリング」画面へ戻ります。「クラスタ数」に「3」と入力した後、「指定せずに分析」のチェックを外します。「出力」の「**クラスタの特徴**」にチェックを入れて、「分析実行」を押してください。クラスタの特徴は各集団の特徴量を出力するためのボタンです。

HAD におけるデンドログラムの作成



階層的クラスター分析を実行すると、新しいシート「**HCluster**」が出現します。分析の結果、集団を 3 つに分割できました。「所属数」には各集団のサンプルサイズ、「所属メンバー」には都市名が記載されています。都市名を確認して、クラスターを解釈しましょう。

クラスターの特徴量は数値と図で出力されます。これらを見れば、クラスター1 は米と中華麺の消費額が多い地域、クラスター2 はパン、生うどん・そば、パスタ、もちの消費額が多い地域、クラスター3 は主食の消費額が少ない地域であることが明らかになりました。

HAD でのクラスター作成結果

